Artificial Intelligence in Safety-critical Systems: A Systematic Review

Yue Wang[1], and Sai-Ho Chung[2]
[1]Centre for Advances in Reliability and Safety, The Hong Kong Polytechnic University, Hong Kong
[2]Department of Industrial and Systems Engineering
The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Corresponding author:
Sai-Ho Chung can be contacted at: nick.sh.chung@polyu.edu.hk

Acknowledgment
This paper (or the work presented in this article) is supported by Centre for Advances in Reliability and Safety (CAiRS), an InnoHK Research Cluster of HKSAR Government.

Abstract

Purpose – This study is a systematic literature review of the application of Artificial Intelligence (AI) in safety-critical systems. The authors aim to present the current application status according to different AI techniques and propose some research directions and insights to promote its wider application.

Design/methodology/approach – A total of 92 articles were selected for this review through a systematic literature review along with a thematic analysis.

Findings – The literature is divided into three themes: interpretable method, explain model behaviour and safe learning. Among AI techniques, the most widely used are Bayesian networks and deep neural networks. In addition, given the huge potential in this field, four future research directions were also proposed.

Practical implications – This study is of vital interest to industry practitioners and regulators in safety-critical domain, as it provided a clear picture of the current status and pointed out that some AI techniques have great application potential. For those are inherently appropriate for use in safety-critical systems, regulators can conduct in-depth studies to validate and encourages their use in the industry.

Originality/value – This is the first review of the application of AI in safety-critical systems in the literature. It marks the first step towards advancing AI in safety-critical domain. The paper has potential values to promote the use of the term “safety-critical” and to improve the phenomenon of literature fragmentation.

Keywords - Artificial Intelligence, machine learning, safety-critical system, neural network, Bayesian, formal verification, adversarial examples

Paper type – Literature review

I. Introduction

The rapid development of Artificial Intelligence (AI) in the past decades has greatly realized its huge potential. Several definitions of AI in [150][151][152][153], primarily refer to building intelligent machines that use computer programs to understand human intelligence and better perform human tasks in ways that are not limited to biologically observable methods. At the same time, machine learning (ML), as the core of data science, is the essence of modern AI, which involves the technology, method, and approach from big data [143][144] to intelligence. AI technology has matured to the point where it can provide practical benefits in many applications [1], including natural language processing [2], computer vision [3] and data mining [4]. Safety-critical system refers to a system failure that can have unacceptable consequences, such as significant loss of life and property or environmental damage [6]. Due to their high requirements for dependability, especially for safety, security and reliability, these systems are widely used in avionics [155], nuclear power plants [156], automotive [157], medical [158] and industrial control systems [159].

In recent years, as the number of safety-critical systems increases, the application of AI in these systems will bring great benefits. Safety-critical systems require extremely high safety requirements as the system failure is a matter of life [164]. Therefore, such systems require certification [6][7] and strong safety guarantees [165]. However, despite its success, AI is not completely reliable, and the accident problem of AI-based systems [5] makes them untrustworthy. For example, cases of AI failure [154], casualties caused by AI-based autonomous vehicles [8][9], and the AI bias for discrimination [166][167]. In addition, most AI-based systems are generally

The following publication Wang, Y. and Chung, S.H. (2022), "Artificial intelligence in safety-critical systems: a systematic review", Industrial Management & Data Systems, Vol. 122 No. 2, pp. 442-470 is published by Emerald and is available at https://doi.org/10.1108/IMDS-07-2021-0419

considered opaque [18] due to their “black-box” nature. An AI “black-box” means that for an AI-based tool, there is no view of how it works for the input and output seen, one example is deep neural network (DNN), which has been criticized for being vulnerable [19]. The risk of trusting “black-box” autonomy algorithms makes AI and ML less acceptable in safety-critical domain [162]. These deficiencies of AI post challenges to the widespread application of AI in safety-critical systems.

In the literature, research on AI safety has been conducted from different angles. From the perspective of ethics: non-maleficence requires safety and security to avoid foreseeable or unintentional harm [12]; ethical governance is the key in building trust [11]. From the perspective of formal methods, AI can be used to design strong, ideally provable and correctness guarantees for mathematically specified requirements [14]. Considering AI safety from a multidisciplinary perspective will provide a basis for understanding that can be transferred to safety-critical systems with a higher level of responsibility [16]. From the perspective of society and economy, the maximization of robustness and beneficial aspects of AI safety research will bring unprecedented benefits to human society [13]. In fact, almost all of these AI safety researches point to the need for AI-based system verification and validation (V&V) in the safety-critical system. In software engineering, V&V is a method consisting of a series of analysis and testing activities that software analysts use to detect bugs or false assumptions in order to gain trust [160]. Conventional V&V approaches for safety-critical systems are testing and model checking [161]. Testing is to use test cases to test the embedded system until it is assumed that all failures have been detected and all tests generated have been passed [17]. Model checking is more through than testing, using an exhaustive search of the system state space to trace logical errors in the specifications [163]. However, both V&V approaches cannot be directly applied to modern safety-critical systems with embedded AI, because such a system is very complex, requires continuous and dynamic interaction with the physical environment. Nevertheless, these V&V approaches generally assume that the system being verified is static, thus the verification is likely to be invalid after the system has learned [13]. Research on “black-box” problem of AI focuses on explainable AI (XAI) and interpretable AI. XAI is defined as a production detail or reason makes the functioning of AI easily understood by the audience [20]. Explainable models can summarize the reasons for neural network behaviour, gain users trust, and generate insights into the reasons for decisions [23]. This academic whirlwind has led toward the concept of responsible AI, further led to legal governance on the explanation of algorithmic decision-making [24]. Interpretability, as the ability to explain or present in a human-understandable manner [23], will be the first step towards creating an explanation mechanism that are necessary for safety-critical tasks [21], especially for high stakes decision-making [22]. It can also confirm other important desiderata of AI systems, such as fairness, reliability, robustness, causality and so on [23]. From a safety perspective, interpretability helps to understand the retrospective and prospective aspects of the AI system [25].

As abovementioned, applying AI to safety-critical systems is fraught with challenges. A comprehensive and in-depth understanding of the current state of applications will allow us to understand why AI is encountering limitations and develop a roadmap to address them. Currently, this field is still in the stage of prudent and exploratory. International standard IEC61508 [26] does not recommend the use of AI technology for electrical, electronic and programmable electronic safety related systems [27]. Besides, the development history of AI is cyclical, it has experienced peaks and troughs, the troughs are called “AI winter”, which is due to overestimation of the ability of AI. The above deficiencies of AI may lead to significant distrust [29] and will limit the AI development process, thereby increasing concerns about the next “AI winter” [28][30][31]. Therefore, to avoid this possible outcome and benefit the safety-critical domain, it is necessary to understand where we are in the development of applying AI in safety-critical systems. This paper aims to provide a holistic overview in the literature, and to give academia, practitioners and regulators a clear picture. To the best of our knowledge, this review is the first in the literature that focuses on the current state of AI development in safety-critical systems.

The remaining part of this paper is organized as follows. Section 2 presents the methodology of the literature review. Section 3 discusses the existing survey paper on ML assurance in safety-critical systems. Section 4 provides the findings including a detailed review of the three themes. Section 5 discusses the future directions on the application of AI in the safety-critical domain and Section 6 concludes the paper.

II. Literature Review Methodology

Systematic literature review is a well-established and rigorous method of evaluating and reviewing research literature based on reproducible, scientific and transparent process, which is a key tool for building an evidence base and reducing bias [10][15]. This review follows the guidelines of [124][168] and consists of four stages: planning, selection, extraction and execution, as shown in Figure 1.

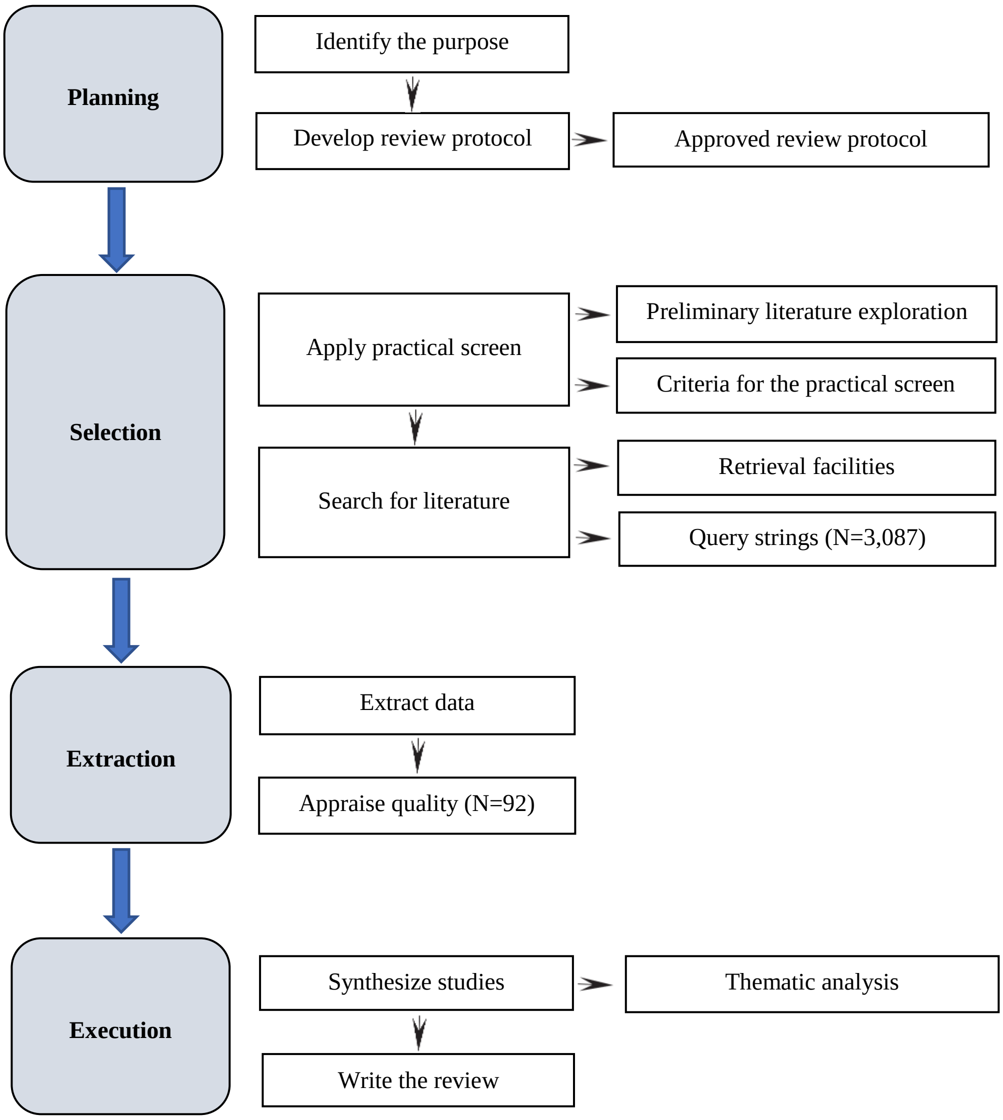


Figure 1 Literature review process

2.1 Planning

The purpose of this study is to systematically review the existing literature, analyze the application status of AI in safety-critical systems, and point out the future research direction. At this stage, prior the start of the review, the review protocol was developed and agreed upon by the authors to carry out the following review.

2.2 Selection

In this stage, the literature was first explored and screened, which provided a preliminary understanding of the literature. Based on the findings in the preliminary literature exploration, we established the practical screen criteria for the following literature retrieval activities. More discussions of this stage are given below.

2.2.1 Apply practical screen

A preliminary literature exploration was carried out on Google Scholar, where the query string was combined with “Artificial Intelligence”, “Machine learning” and “Safety-critical system”. Some of the findings are: (1) fewer articles are obtained from searching strings “Artificial Intelligence” “Safety-critical system”, and most do not directly correspond to the intent of this paper. Instead, more articles can be found by “Machine learning” “Safety-critical system”, but most papers are about specific ML and AI techniques, such as neural networks and Bayesian networks (BNs). (2) Some articles concerning process plants, process systems, subsea blowout preventer control systems are safety-critical, but there is no mention of “safety-critical” either in the article or in the title; instead, some use “critical system”, “system safety” or even just “system”. (3) Some articles do not explicitly mention “safety-critical” but focus on the “dependability” of safety-critical systems and its basic attributes – “safety” and “reliability”. (4) A significant number of ML articles on safety-critical systems and applications are conference papers and conference proceedings. (5) Some of the relatively highly cited articles were preprints from arXiv. In computer science, arXiv is a free public server repository for electronic e-prints of research, which is hosted by Cornell University [122].

Based on the above findings, the authors recognize that the literature on the use of AI in safety-critical systems is fragmented. To cover a more comprehensive and objective perspective, some practical screening criteria have been developed: (1) this paper does not limit the sources of literature to specific publishers, journals or conferences, but covers a variety of scientific research media, including major conferences, journal articles sponsored by different publishers, and widely cited arXiv preprints were also included through careful review of relevance, citations, and quality. (2) Papers selected in this review cover a time span of the last three decades in order to have a comprehensive coverage. (3) The content of the paper must be relevant to the purpose of this review. (4) Articles that match the query string but do not meet the objective of this review, work-in-progress reports, as well as articles that are not in English were discarded. (5) The academic fields were limited to computer science, engineering and technology, mathematics and systems science.

2.2.2 Search for literature

During paper identification, Google Scholar and the Web of Science were selected as retrieval facilities. Table 1 shows all the query strings in three categories: Category 1 are all AI-related, Category 2 are attributes related to safety-critical systems, and Category 3 are various descriptions of safety-critical systems. The search sentence pattern takes the form of a random iterative combination of Category 1 plus Category 2 plus Category 3, with one query string from each category at a time (for example, “Machine learning” “Safety” “Critical System”). The sentences pattern search iterates until most of the combination forms are combined. After this process, a total of 3,087 papers were included.

No.	*Category 1*	*Category 2*	*Category 3*
1	Artificial Intelligence	Dependability	Safety-critical system
2	Machine learning	Safety	Safety-critical
3	Neural network	Reliability	Critical system
4	Bayesian	Verification & Validation	System safety

Table 1 Query strings during paper identification

2.3 Extraction

For each article found in the previous process, the title, abstract, introduction and conclusion sections were read by the authors to determine their relevance, followed by another round of quality checking. Articles that are completely related to computer science but do not about AI application to safety-critical systems were discarded. After this process, 92 articles were selected in this review for intensive reading. Among them, there were 36 journal papers, 44 conference proceedings papers, 11 arXiv preprints, and one technical report paper. As shown in Figure 2, the application of AI in safety-critical systems is a promising research field and has received great attention since 2015.

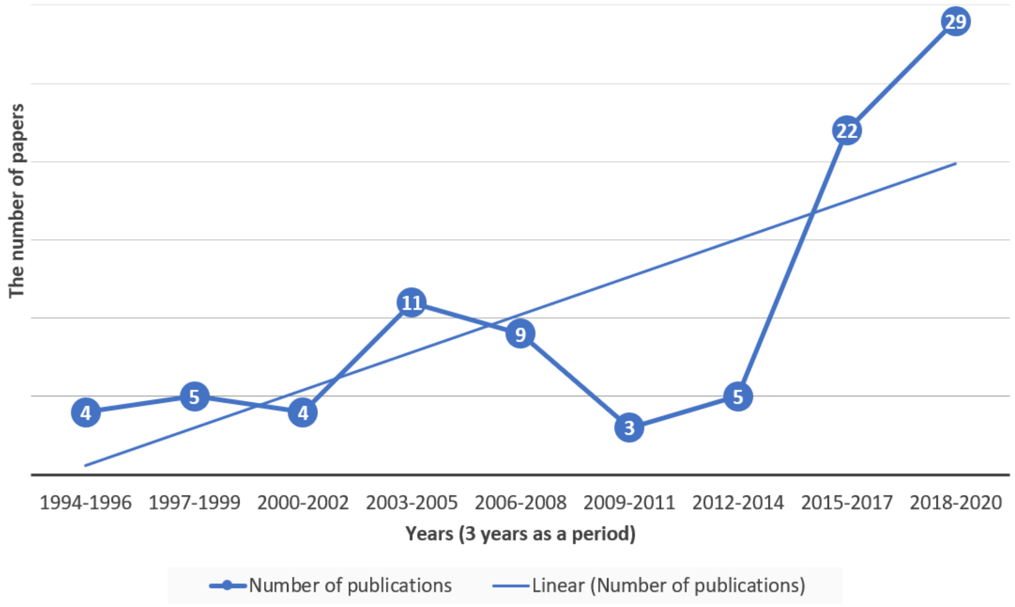


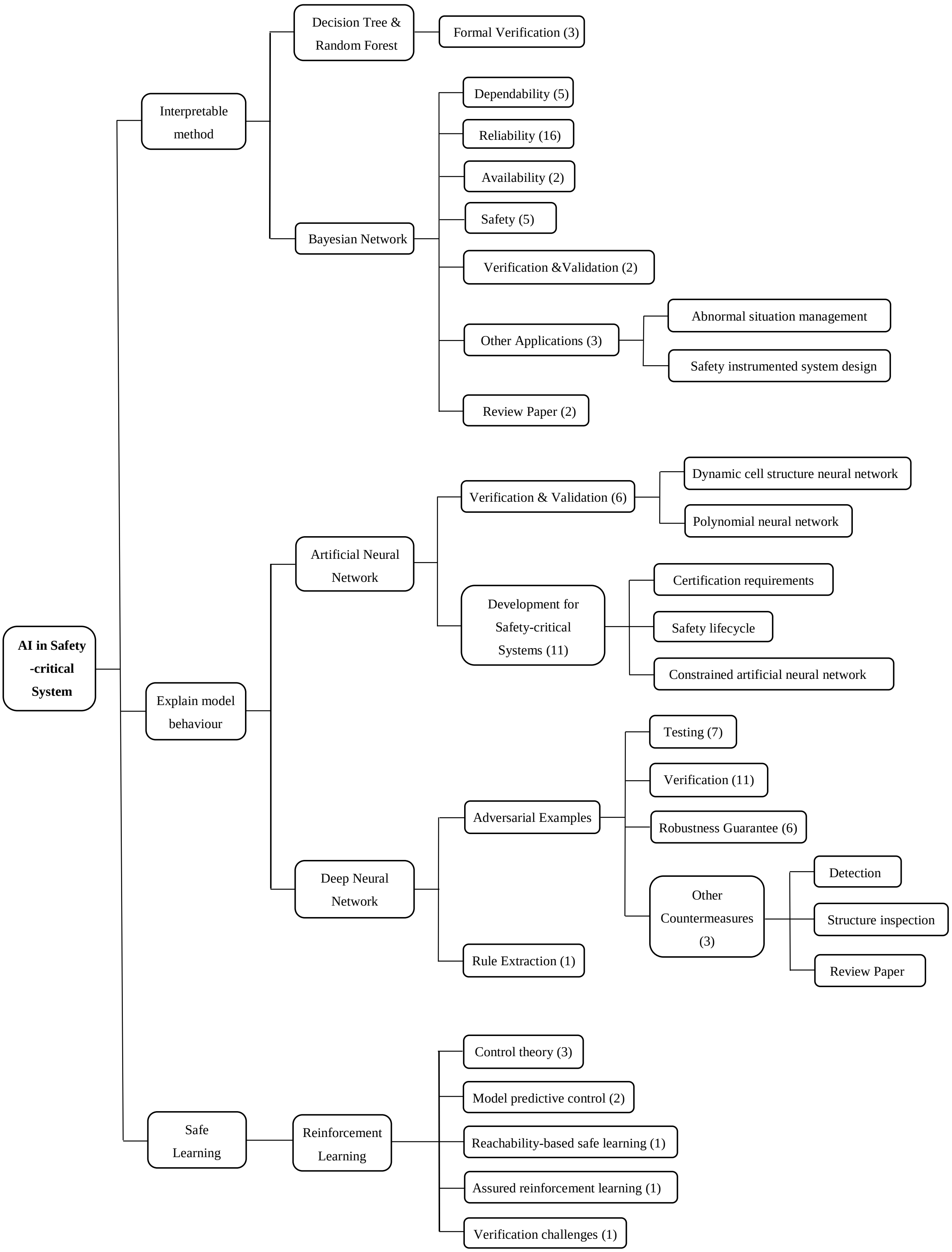
Figure 2 Distribution of papers over time

2.4 Execution

In the process of research synthesis, thematic analysis [145] was adopted, which is a qualitative context analysis method to identify themes. There are five steps to this process: (1) familiarize with the data by reading; (2) decompose the data into small chunks and generate initial codes; (3) through examining the codes and search for broader themes characterized by its significance; (4) review, modify and develop preliminary themes; (5) define themes. For the 92 papers selected in the extraction stage, the authors summarized the content of each paper. Among them, there is one survey paper on ML assurance in safety-critical systems, which is discussed in Section 3. The authors then extracted small chunks from the remaining 91 papers as initial codes and classified them according to five AI techniques: decision tree and random forest, Bayesian network, artificial neural network, deep neural network and reinforcement learning. After a careful study of the codes and the contents of the article on these five AI techniques, the authors realized that papers on the use of decision tree and random forest as well as Bayesian network in safety-critical systems focused on interpretable methods, as these AI techniques are transparent in nature. However, due to the "black-box" problem of neural networks, papers on artificial neural network and deep neural network are focused on explaining their model behaviour. In addition, papers on reinforcement learning are focused on safe learning. On this basis, three themes were determined: interpretable method; explain model behaviour and safe learning. Figure 3 shows the three themes and the chunks they contain, including the number of articles reviewed in this paper. These three themes are discussed in detail in Section 4.

III. Survey Paper on Machine Learning Assurance in Safety-critical Systems

Ashmore *et al*. [121] conducted the first comprehensive and state-of-the-art survey of safety-critical systems from a ML assurance perspective, covering the desiderate, methods and challenges to achieve such assurance across the entire ML lifecycle. The data management phase must ensure the relevance, completeness, balanced and accurate datasets. During model learning, interpretable ML models are critical for safety-critical systems to aid assurance and to ensure models are reusable and interpretable by providing evidence. Model verification is essential, and formal verification can be used to determine the suitability of ML models before integrated into safety-critical systems. During model deployment, the broader system must be able to tolerate the occasional erroneous output of ML models and its output must be explainable. Extending the applicability of ML to safety-critical applications required a higher level of assurance than current ML applications. The key is to generate evidence that is fit for purpose and can be fully integrated into the system to gain trust and demonstrate the safety of ML component.


AI in Safety -critical System
Interpretable method
Decision Tree & Random Forest
Formal Verification (3)
Bayesian Network
Dependability (5)
Reliability (16)
Availability (2)
Safety (5)
Verification &Validation (2)
Other Applications (3)
Abnormal situation management
Safety instrumented system design
Review Paper (2)
Explain model behaviour
Artificial Neural Network
Verification & Validation (6)
Dynamic cell structure neural network
Polynomial neural network
Development for Safety-critical Systems (11)
Certification requirements
Safety lifecycle
Constrained artificial neural network
Deep Neural Network
Adversarial Examples
Testing (7)
Verification (11)
Robustness Guarantee (6)
Other Countermeasures (3)
Detection
Structure inspection
Review Paper
Rule Extraction (1)
Safe Learning
Reinforcement Learning
Control theory (3)
Model predictive control (2)
Reachability-based safe learning (1)
Assured reinforcement learning (1)
Verification challenges (1)

* Note: The numbers in brackets indicate the numbers of articles reviewed by this paper under that category

Figure 3 Three themes and chunks in the literature

IV. Findings

This section consists of four parts. Section 4.1 explains the literature reviewed. Section 4.2 to 4.4 discuss the literature on interpretable method; explain model behaviour and safe learning, respectively.

4.1 Distribution of Literature

Figure 4 shows the application of different AI techniques in the literature over time. The past decade has seen an explosion in the application of ML in this field, particularly DNNs. The authors found that research on DNN mostly focused on image classification in safety-critical applications, which deserves to be discussed separately. Therefore, DNN and Artificial neural network (ANN) are classified respectively in this paper. In addition to DNNs, the application of ANNs in safety-critical systems has seen a significant decline in attention in the past decade, whereas other ML algorithms and techniques such as reinforcement learning, decision tree and random forest have emerged. In contrast, the application development of the BN in safety-critical systems is relatively stable. All the articles selected in this review paper are shown in Table 2.

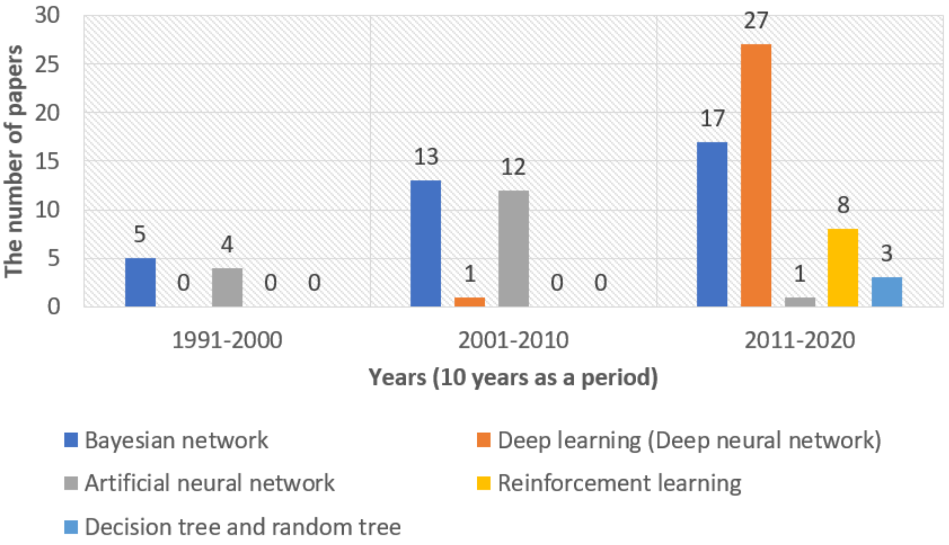


Figure 4 Distribution of papers on AI techniques over time

4.2 Interpretable Method

Interpretable AI is an algorithm or model that can clearly explain its decision-making process. Interpretable method is further divided into two categories, Decision Tree & Random Forest and BN. Both of these AI techniques are transparent in nature [32]. At present, the application of decision tree in safety-critical systems is still in its preliminary stage, and the focus of the literature is on formal verification. Unlike decision trees, BNs have a relatively long history of application in system safety and reliability, most of the research focuses on system analysis.

4.2.1 Decision Tree and Random Forest

Decision tree is a ML algorithm which is “naturally” explainable, the input data can be clearly traced passing through the model. It has been widely used in exploratory data analysis and predictive modelling applications [123]. Random forest is used to solve the overfitting problem of decision trees, it adopts the idea of ensemble learning, which integrates multiple random decision trees together and aggregates their predictions by means of averaging. It shows excellent performance in settings where the number of variables is far greater than the observations and extends the ability of decision trees to solve large-scale problems [125].

Papers	No. of papers	Classification
Formal Verification: Törnblom and Nadjm-Tehrani (2018) [33]; Törnblom and Nadjm-Tehrani (2019) [35]; Törnblom and Nadjm-Tehrani (2019) [34]	3	Decision Tree & Random Forest
Dependability: Neil *et al.* (1996) [55]; Fenton *et al.* (1998) [56]; Kang and Golay (1999) [57]; Bobbio *et al.* (2001) [58]; Montani *et al.* (2005) [59] Reliability: (Reliability Analysis) Torres-Toledano and Sucar (1998) [36]; Langseth and Portinale (2007) [37]; Simon *et al.* (2007) [39]; Montani *et al.* (2008) [40]; Simon *et al.* (2008) [38]; Cai *et al.* (2012) [46]; (Reliability Modelling) Weber and Jouffe (2003) [42]; Boudali and Dugan (2005) [44]; Weber and Jouffe (2006) [43]; Doguc and Ramirez-Marquez (2009) [45]; Amrin *et al.* (2018) [41]; Liu and Liu (2019) [47]; (Fault Diagnosis) Chiremsel *et al.* (2016)[148]; Yang *et al.* (2008) [48]; Cai *et al.* (2017) [49]; Cai *et al.* (2017) [146] Availability: Amin *et al.* (2018) [60]; Wang *et al.* (2020) [147] Safety: Bouissou *et al.* (1999) [61]; Gran (2002) [53]; Khakzad *et al.* (2011) [50]; Khakzad *et al.* (2013) [51]; Prabhakaran *et al.* (2016) [54] Verification & Validation: Schietekat *et al.* (2016) [149]; Douthwaite and Kelly (2017) [64] Other applications: Kannan (2007) [65]; Naderpour *et al.* (2014) [62]; Naderpour *et al.* (2015) [63] Review Paper: Weber *et al.* (2012) [66]; Kabir and Papadopoulos (2019) [67]	35	Bayesian Network
Verification & Validation: Andrews *et al.* (1995) [81]; Hull *et al.* (2002) [83]; Taylor *et al.* (2003) [80]; Darrah *et al.* (2004) [85]; Taylor and Darrah (2005) [82]; Darrah *et al.* (2005) [84] Development for Safety-critical Systems: Certification requirements: Bedford *et al.* (1996) [78] Safety lifecycle: Rodvold (1999) [71]; Weaver *et al.* (2002) [72]; Kurd and Kelly (2003) [68]; Kurd and Kelly (2003) [73]; Kelly (2004) [70]; Kurd and Kelly (2005) [77]; Kurd *et al.* (2007) [69]; Ward and Habli (2020) [74] Constraint artificial neural network: Wen *et al.* (1996) [75]; Kurd and Kelly (2007) [76]	17	Artificial Neural Network
Adversarial Examples: Testing: Pei *et al.* (2017) [88]; Tian *et al.* (2018) [9]; Ma *et al.* (2018) [86]; Ma *et al.* (2018) [87]; Sun *et al.* (2018) [89]; Sun *et al.* (2018) [90]; Zhang *et al.* (2020) [92] Verification: Pulina and Tacchella (2010) [95]; Huang *et al.* (2017) [93]; Pei *et al.* (2017) [103]; Xiang *et al.* (2017) [99]; Katz *et al.* (2017) [94]; Cheng *et al.* (2017) [96]; Carlini *et al.* (2017) [100]; Bunel *et al.* (2018) [97]; Wang *et al.* (2018) [98]; Xiang *et al.* (2018) [101]; Wicker *et al.* (2018) [91] Robustness Guarantee: Hein and Andriushchenko (2017) [108]; Gopinath *et al.* (2017) [109]; Gehr *et al.* (2018) [102]; Hendrycks and Dietterich (2019) [105]; Hein *et al.* (2019) [106]; Croce *et al.* (2019) [107] Other countermeasures: Wang *et al.* (2019) [104]; Papernot and McDaniel (2018) [110]; Yuan *et al.* (2019) [111] Rule Extraction: Hailesilassie (2016) [112]	28	Deep Neural Network
Control theory: Berkenkamp *et al.* (2016) [115]; Berkenkamp *et al.* (2017) [114]; Richards *et al.* (2018) [113] Model predictive control: Kahn *et al.* (2017) [120]; Koller *et al.* (2018) [116] Reachability-based safe learning: Akametalu *et al.* (2014) [117] Assured reinforcement learning: Mason *et al.* (2017) [119] Verification challenges: Wesel and Goodloe (2017) [118]	8	Reinforcement Learning

* Note: The underlined words in the table represent the third level chunks in Figure 3

Table 2 Papers reviewed in this study

4.2.1.1 Formal Verification

One of the constraints of AI utilizing ML algorithms in safety-critical systems is the lack of verification methods and the difficulty of interpretation with large datasets. Formal verification can be adopted in the early process of system development, which provides methods and techniques for mathematically proving the correctness of a system [126], it has mainly used in safety-critical domains, such as military [128] and aerospace [127]. The functional safety standard IEC61508 allows for the first time the use of formal verification methods during the certification process [129]. However, when verifiability is critical, decision trees are more appropriate to address this challenge in terms of their simplicity. In addition, decision trees and random forests are easier to analyze systematically. However, a major limitation is combinatorial explosions in large models.

Törnblom and Nadjm-Tehrani [33] first proposed a formal method to verify the properties of random forest, which divides the input domain of the decision tree into disjoint sets and explores all path combinations to offset the combinatorial path explosion. The method is implemented through an automated computing tool for enumerating and verifying equivalence classes. Further, Törnblom and Nadjm-Tehrani [35] generalized the above method to gradient boosting machines and proposed a formal verification method of tree ensembles implemented through an update tool. It extracts equivalence classes from decision trees and tree ensembles, and uses formal verification to prove that their input-output mappings met the requirements. However, the above two methods struggle with combinatorial explosions on high-dimensional data. Thus, Törnblom and Nadjm-Tehrani [34] proposed a formal iterative abstraction-refinement method, which made it possible to formal verify tree ensemble for high-dimensional data training. The results showed that the performance of the improved by several orders of magnitude.

4.2.2 Bayesian Network

BN was originated in the field of AI and has become a robust and effective framework for uncertain knowledge reasoning. To improve safety, it is necessary to learn from the past, BN has the features of updating, inference and diagnosis, it can reduce uncertainty so as to improve the cognition and understanding of the complex system [130]. As a powerful decision supporting tool, BNs have been widely used in the practical applications of predicting the performance of safety-critical systems [131]. Figure 5 summarizes the development timeline of the various applications of BNs in safety-critical systems.

4.2.2.1 Dependability

Traditional approach to dependability assessment relies on expert judgement, however, this is generally considered to be uncertain and an ad-hoc procedure. Neil *et al.* [55] were among the earliest advocates of the transformation from ad-hoc assessment to argumentation with the usage of BN. Later, a project named DATUM presented in Fenton *et al.* [56] was the first application of BNs in the dependability assessment of software-intensive safety-critical system. The methods and techniques of uncertainty modelling were firstly studied in-depth and BN was selected as the most appropriate modelling method. Kang and Golay [57] first applied BN to dependability analysis of a complex nuclear power plant through a BN-based diagnostic advisory system framework to improve its operational availability. In addition, some studies have compared BNs to fault trees, which is one of the most commonly used and popular techniques for dependability analysis of large-scale safety-critical systems. Bobbio *et al.* [58] proposed a conversion algorithm to convert the fault tree into BN and further pointed out that any fault tree can mapped to BN directly. Dynamic fault tree (DFT) improves fault tree modelling power by introducing new primitive gates that can accommodate complex dependencies. Dynamic Bayesian network (DBN) adopts a discrete time approach and has the advantage of providing a unified framework in which both static and dynamic components can be analyzed. Montani *et al.* [59] provided a translation of DFT into corresponding DBN, by characterizing dynamic gates in the DBN framework.

4.2.2.2 Reliability

Reliability is one of the most important dependability attributes in systems engineering, it refers to the probability that equipment satisfactorily performs its expected function and has no failure within a certain mission time under specific design and environmental conditions. Safety-critical systems of complex industrial plants and critical application equipment require high reliability, but it is almost impossible to model the entire system as these systems becoming more complex. An early study by Torres-Toledano and Sucar [36] suggested a BN-based computational method to conduct reliability analysis by explicitly clarifying dependency between failure, including the effects of maintenance. It has also presented a general methodology to reliability modelling of complex systems. Langseth and Portinale [37] comprehensively discussed the applicability and proposed BN as a reliability analysis framework. One of the key attractions of BN is its ability to combine information from different sources and provide a global safety assessment.

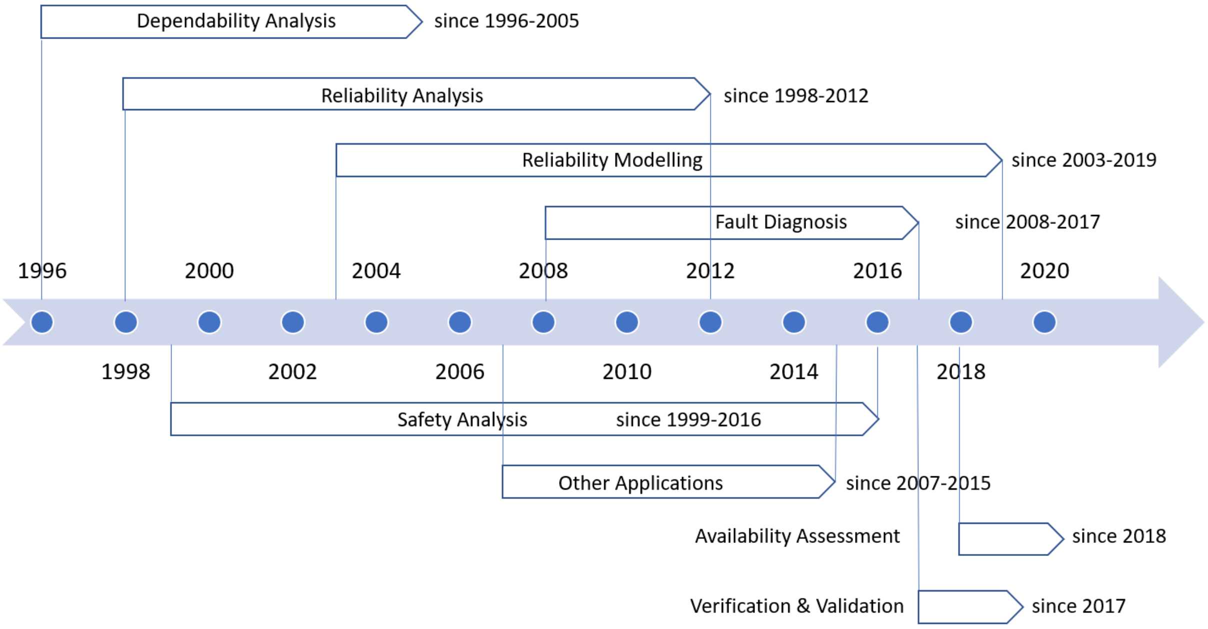


Figure 5 Timeline of applications of Bayesian networks for safety-critical systems

Some studies have proposed to combine BN with evidence theory for under epistemic uncertainty, which resolves incomplete data in reliability context and inconsistency between system model and reliability model. Simon *et al.* [39] proposed an evidential network generation method integrating evidence theory with the BN, during the allocation process, new attributes can be obtained to manage epistemic uncertainty in reliability analysis. Later, Simon *et al.* [38] proposed the use of uncensored data and applied the Dempster Shafer theory to BN tools to extract as much information as possible from the available data. Furthermore, Montani *et al.* [40] extended reliability analysis to the use of DBN and proposed a software tool RADBAN which implements a modular algorithm automatically translating a DFT into its corresponding DBN and deduces it using the classical algorithm. BN application in reliability evaluation of safety-critical system was presented by Cai *et al.* [46], it has been applied to reliability evaluation of redundant system of a subsea blowout preventer (BOP) control system.

Markov chain used to be a popular tool to model the reliability of systems, but it encounters state combination explosions as system become complex. Weber and Jouffe [42] introduced DBN as the equivalent model of the Markov chain and further present a methodology for developing DBNs to formalise complex dynamic models. A novel reliability analysis and modelling framework based on discrete time BN is proposed by Boudali and Dugan [44] to cope with the increasingly complex systems. In addition, using BNs to model systems with many variables often results in complex models, but object oriented Bayesian networks (OOBNs) are ideal and useful for the process modelling of industrial systems. Weber and Jouffe [43] proposed a powerful tool for maintenance decision-making based on the dynamic object oriented Bayesian networks (DOOBNs) model. The model decomposes the global network into a hierarchical structure and can easily model the temporal behaviour of complex system state probability by designing a DOOBN structure.

However, the use of BN to estimate the reliability of system must be known a priori, which relies heavily on experts. Such experts are limited and are not always available, and human intervention may lead discrepancies. The first study by Doguc and Ramirez-Marquez [45] introduced the automatic construction of a BN model for estimating the system reliability without the involvement of human experts. In addition, there is no defined semantics to guide the model formation of BN in reliability modelling, one solution is to use semantic method. Amrin *et al*. [41] proposed a novel method through an “Idea Algebra” framework automatically generates BN for reliability analysis directly from the system description. A method of constructing an equivalent BN based on GO model was proposed by Liu and Liu [47], which has been further applied to reliability assessment of a subsea BOP control system.

With the rapid development of modern industrial systems, fault diagnosis must be fully leveraged to quickly detect process anomalies and component faults to locate the root cause of failures. BN is a powerful risk analysis tool for fault diagnosis through backward analysis. Chiremsel *et al*. [148] proposed a hybrid probabilistic fault diagnosis method for safety instrumented systems in the oil and gas industry based on fault tree analysis and the equivalent BN. Cai *et al*. [49] reviewed the application of BNs in fault diagnosis of engineering systems in the past decade with the general procedure of fault diagnosis modelling. However, one disadvantage of BN is it requires too much prior probability information so it is beneficial to combine fuzzy logic with Bayesian reasoning. Cai *et al*. [146] proposed a DBN-based fault diagnosis method for transient and intermittent faults of industrial safety-critical systems. This method takes into account the dynamic behaviour of the system and identify fault components and distinguish fault types. Yang *et al*. [48] proposed a Failure mode and effects analysis (FMEA) tool that is fuzzy rule-based Bayesian reasoning, which can support safety decision making in the case of subjective data.

4.2.2.3 Availability

Availability refers to the ability to operate and maintain an item in a specified manner to perform its specific functions within a given amount of time. Amin *et al*. [60] proposed a dynamic availability assessment technique based on DBN and further applied to two safety-critical systems. This methodology takes into account the dependence and independence among the failure cause factors and can help identify the most critical failure causes. Wang *et al*. [147] established the fault tree of subsea Xmas tree system used in offshore oil and gas development under fault mode and transformed it into DBN and further analyzed the reliability and availability of subsea tree system under different repair states.

4.2.2.4 Safety

The concept of safety software development has been put forward along with the increasing applications of programmable devices in safety-critical systems. However, the characteristics of software make it difficult to assess its reliability as the preliminary failure is usually due to design fault which is difficult to predict. Therefore, the assessment of a software-based safety-critical system can only be qualitative. One way to license such a system is to build a “safety case” or “safety argument”, a collection of various pieces of evidence related to the development process and the final product. The SERENE European project was aimed at building a method and tools that can facilitate the availability of such high-level safety argument to improve the repeatability and ease of understanding of safety assessments. Bouissou *et al*. [61] presented the result of this project performed at Électricité de France to helped assessors building safety arguments by a BN safety assessment model appropriately weighing various sources of evidence to arrive at a final judgement. On this basis, Gran [53] combined BN and software safety standard (DO-178B) to conduct safety assessment of the software-based system.

Proper operation of safety-critical systems is critical, safety analysis of safety-critical process facilities can ensure safety and reliability. Khakzad *et al.* [50] comparing fault tree with BN in process facility safety analysis and proved that BN is superior and has a flexible structure that can be adopted to a variety of accident scenarios. Bow-tie as a popular process system safety analysis technique cannot be used in dynamic safety analysis due to its static structure. Khakzad *et al*. [51] presented a dynamic safety analysis method by mapping the bow-tie to BN. Prabhakaran *et al*. [54] proposed a safety assessment approach for an unmanned aerial vehicle (UAV) safety-critical system to appropriate monitoring the safety-critical outputs. This enabled the use of BNs to secure the system and to ensure its fully successful performance.

4.2.2.5 Verification & Validation

Assurance of safety-critical autonomous systems and their driving technologies is a major research challenge. Schietekat *et al*. [149] proposed a V&V methodology for an BN-based aircraft vulnerability system. It first defines the reality, then develops, computerized and exercised a BN-based conceptual model. Throughout the process, V&V continues at each step and evidence is recorded. However, this study is conceptual in nature as it has not been applied to real-world models. The first step to develop a rigorous V&V approach was presented by Douthwaite and Kelly [64], the study proposed a reference model to support the comprehensive description and modelling of BN-based safety-critical systems and further proposed a method for developing generic and system-specific V&V objectives.

4.2.2.6 Other Applications

Kannan [65] discussed the suitability of using BN as a “live” model during the design phase of safety instrumented systems. For human operations in safety-critical systems, situation awareness (SA) is the key factor to improve performance and reduce errors. Naderpour *et al*. [62] proposed the first situational network modelling process with the establishment of a situational assessment model based on DBN and risk indicator. On this basis, Naderpour *et al*. [63] further proposed an abnormal situation modelling method with the specific capabilities of BNs and fuzzy logic systems to determine abnormal situations.

4.2.2.7 Review Paper

BNs have been widely used in the area of system safety and reliability. Weber *et al*. [66] conducted a bibliographical review of 200 specific literatures on applications in dependability, risk analysis and maintenance. The study concluded that the use of BN is a trend due to its benefits compared with other classical methods, but its shortcoming lies in the lack of concrete semantics to guide model development to ensure coherence. The authors further advocated to transform the classical dependability model into BN and to define a new model development method. This review covered a wide range on complex industrial systems. Another comprehensive review was conducted by Kabir and Papadopoulos [67] on the use of BNs and Petri nets to assess safety, reliability, and risks. The authors believe that these attributes of the system ensure the dependability of safety-critical systems must be performed throughout the lifecycle of the system.

4.3 Explain Model Behaviour

Explaining model behaviour focuses on neural networks, subdivided into the ANN and DNN. ANNs have been criticized as “black-box” which makes them difficult to conduct system analysis and defect detection for safety-critical applications. The current research direction is mainly focused on ways to explain its behaviour. Research on V&V shows that rule extraction can extend ANN to safety-critical systems, another research direction on the construction of a hybrid ANN with transparency and interpretability. On the other hand, DNNs have been widely used in computer vision-based safety-critical autonomous systems, such as image classification, but a common problem is that DNNs are considered vulnerable in the face of adversarial examples.

4.3.1Artificial Neural Network

ANNs are highly parameterized nonlinear models composed of processing neurons, which can be used to approximate the relationship between the input and output signal of complex systems [52]. It can approximate any continuous function, but the immediate structure of the fitting model prevents it from providing insight into the relative importance, underlying relationship, and model structure to the outcome [132]. Developers of ANN have been cautious in extending it to applications in safety-critical systems [80], where the reliability requirements takes precedence over capability [133].

4.3.1.1 Verification & Validation

Explanation capability is essential for ANN-based safety-critical applications. However, the lack of explanation capability limits the full realization of such ANN-based systems. Rule extraction from trained ANNs can provides user acceptance and extend such systems to safety-critical applications. It is a formal method that transforms a “black-box” system into a “white-box” system by translating the internal knowledge into a series of symbolic rules, which can be further used for V&V and certification of ANNs. A survey conducted by Andrews *et al*. [81] reviewed various algorithms for rule extraction and proposed a new classification scheme: decompositional, pedagogical and eclectic. Taylor and Darrah [82] identified several areas in which rule extraction can be used for the V&V of ANN. The authors considered rule initiation and rule insertion, as two methods of rule extraction, both of which can be used in V&V throughout the development lifecycle.

Traditional training-validation-testing approach cannot assure ANN meet the rigorous requirements for safety-critical systems. Taylor *et al*. [80] discussed the trend and potential usefulness techniques for V&V of ANN, including improved testing, formal methods, run-time monitoring, cross validation, and visualization. The conclusion is that different methods can be applied to different stages of the ANN lifecycle, but there is a lack of methodology to provide V&V practitioners with the assurance of ANNs in safety-critical systems.

There are some practical studies on validation of flight-critical systems. Hull *et al*. [83] proposed an analysis technique to replace the lookup table in various safety-critical control applications. It is based on using Lipschitz constants to offer guaranteed bounds and can be used as part of a polynomial neural networks (PNNs) verification procedure. For adaptive neural networks, an algorithm for extracting rules from dynamic cell structure (DCS) neural networks is proposed by Darrah *et al*. [85]. These rules extracted can be used to assist the V&V of neural networks in safety-critical applications. On this basis, a geometric algorithm for extracting deterministic rules is further proposed by Darrah *et al*. [84], which can extract the rules that are consistent with the neural network.

4.3.1.2 Developing ANNs for Safety-critical Systems

The application of ANNs in safety-critical systems is typically limited to advisory roles and do not have the final say in decision-making. It is important to conduct certification process prior to its application. Bedford *et al*. [78] discussed the requirements for a standard certification. The authors argued that the key problem with ANNs is its inability to analyze its behavior in a “white-box” way, due to the lack of a compelling safety argumentation. In the literature on safety-critical systems, there are two types of safety arguments: process-based and evidence-based. Rodvold [71] proposed a nested loop model of the software development process for process-based safety argumentation, which is specifically used for the network development of ANNs in safety-critical applications. Kelly [70] proposed a systematic safety case development method based on goal structuring notation (GSN), which is a graphical argumentation notation clearly and explicitly represent the various elements of a safety argument and their relationships. Process-based argument can be made through the proposed software development process. Weaver *et al*. [72] proposed an evidence-based framework for generating software product-based safety arguments represented by GSN. Product-based argument of the functional behavior of ANN is obtained by meeting safety criteria. Kurd and Kelly [68] and Kurd *et al*. [69] defined the minimum behavioral properties that must be enforced in safety-critical applications from a high-level perspective to generate potential "white-box" analysis. Kurd *et al*. [69] and Kurd and Kelly [73] proposed the safety lifecycle of artificial neural networks (SLANN) in safety-critical applications, which is a hybrid ANN combines symbolic and neural network paradigms of the "W" model. It can manage the behavior represented by ANN and providing acceptable form of safety assurance. This approach has great potential for providing "white-box" analysis through rule-extraction algorithms and offers the possibility of analyze using decomposition methods. When applying ML algorithms to safety-critical systems, interpretability is critical to understand how the algorithm works. Ward and Habli [74] proposed an argument pattern expressed in GSN to prove the sufficient interpretability of ML models in a wider assurance case.

Some research into the combination of ANN with fuzzy logic to achieve "white-box" analysis and transparency for use in safety-critical applications. Wen *et al*. [75] proposed an iterative Neuralware engineering framework to maintain consistency between specification, model checking, and formal testing. It based on the design of a fuzzy neural network, which is a "hybrid" of ANN and fuzzy logic. Kurd and Kelly [76] proposed a "neuro-fuzzy" model named the safety critical artificial neural network (SCANN). It is a "hybrid" of ANN and fuzzy self-organizing map (FSOM) and can translates fuzzy rules into SCANN by inserting rules without affecting the fidelity. In order to test the practicality of SLANN and SCANN, Kurd and Kelly [77] evaluated their practicality in a real-world problem of Gas Turbine Aero-Engine Control. The SLANN based on decomposition and analytical approach provided the feasibility and effectiveness in the safety of the development process. Using safety constraints, the complete behavior of SCANN can be easily extracted and controlled. Both models demonstrate that the use of neural networks and fuzzy logic systems in safety-critical applications presents evidence-based safety arguments.

4.3.2 Deep Neural Network

DNN-based systems are increasingly being used in safety-critical autonomous systems, where the response behavior of the system to corner-case input is particularly important [88]. However, DNNs are thought to be vulnerable to adversarial examples by slightly perturbing the original examples [79]. In addition, adversarial examples with transferability also expose weakness in system robustness, posing a serious challenge [134]. Existing research to address this problem is mainly in providing testing, formal verification, robustness guarantee and other countermeasures. Figure 6 summarizes the development timeline of the various applications of DNNs in safety-critical systems.

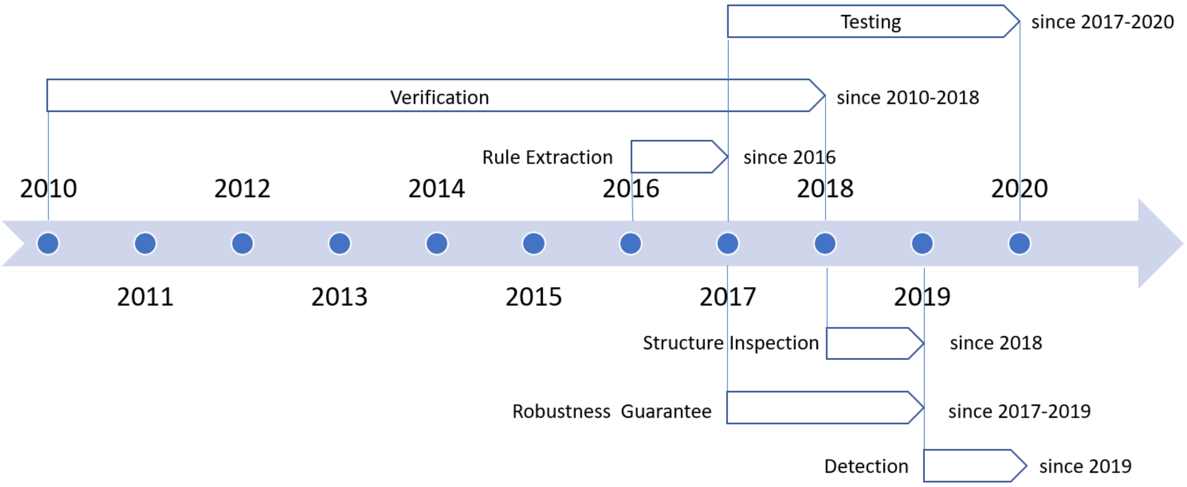


Figure 6 Timeline of applications of Deep neural networks for safety-critical systems

4.3.2.1 Adversarial examples

a. Testing

A DNN-based safety-critical system must be systematically tested. Traditional testing approach is to divide the training and testing set randomly, thus the quality of the test set is key in gaining trust. However, in this case, DNNs only perform a portion of all rule learning, and it is difficult to build a robust safety-critical system using such a test set alone. For DNN testing, there has been some research on neuron coverage. Pei *et al*. [88] proposed DeepXplore, the first "white-box" framework to systematically test real-world DNN systems by introducing neuron coverage as a "white-box" testing metric allows automatic identification of erroneous behavior without the need for manual labeling. Tian *et al*. [9] proposed DeepTest, the first systematic and automated testing tool that systematically test the safety-critical erroneous behavior of DNN-based autonomous vehicles without providing any theoretical guarantee. Some studies have discussed the testing criteria for DNN systems. Ma *et al*. [86] proposed DeepGauge, a set of multi-granularity testing criteria inspired by the traditional MC/DC software testing criterion. Sun *et al*. [89] proposed a "white-box" testing method consisting of four testing criteria. This was the first study to capture and quantify the causal relationship that exists in the DNN. The above four studies are referred to as concrete execution. Sun *et al*. [90] proposed DeepConcolic, the first concolic testing method and a hybrid software testing technique. It combines concrete execution with symbolic analysis. Mutation testing as a relatively mature testing technique in traditional software testing can systematically evaluate software quality and locate defects. Inspired by this, Ma *et al*. [87] proposed DeepMutation, the first mutation testing framework used to measure the quality of test data. In addition, Zhang *et al*. [92] conducted a comprehensive survey of 144 ML testing papers in different published fields and suggested future research directions of ML testing.

b. Verification

It is important to thoroughly validate safety assurance of the DNN output behavior. In the literature, studies have focused on formal verification. Researchers have attempted to extend and customize a theoretical solver in the context of Satisfiability Modulo Theory (SMT) to estimate decision boundaries with strong guarantees. Pulina and Tacchella [95] performed the first formal verification and proposed an abstraction-refinement method that verifies the safety of Multi-Layer Perceptrons (MLPs) by abstracting linear arithmetic constraints into Boolean combinations. Huang *et al.* [93] proposed a general framework for automatic safety verification of DNN classification decisions and proven the robustness of local adversarial and detect misclassification. Katz *et al.* [94] proposed a simplex based Reluplex algorithm to verify the properties of DNN containing a linear function and a Rectified linear unit (ReLU) activation function. The ReLU activation function has the advantages of faster training process avoiding gradient varnishing, and the piecewise linearity of the ReLU activation function allows the DNN to be generalized to previously unseen inputs. Xiang *et al.* [99] proposed a formal verification method for MLPs with ReLU activation function, the authors use a layer-by-layer approach to compute the output of the reachable set for safety verification, which relies on the ReLU functions reachability analysis in the form of a set of manipulations on the joint of polytopes. Xiang *et al.* [101] proposed another method for computing a reachable set based on simulation and develop automatic safety verification. Instead of targeting specific activation functions, the study formulated the problem as a set of optimization problems. Cheng *et al.* [96] proposed a formal verification approach by defining resilience properties of a tolerable DNN-based classifier based on mixed integer programming (MIP) to compute the maximum perturbation bound. Bunel *et al.* [97] proposed a unified framework and new benchmark dataset that contains a collection of previously published testcases. Carlini *et al.* [100] proposed a method of using formal verification to evaluate the effectiveness of an attack and defense against the adversarial example. The key idea is to use verification to construct an adversarial example with provably minimally distortion. Wang *et al.* [98] proposed an effective method for formal safety analysis based on symbolic linear relaxation and directed constraint refinement.

There are some "black-box" verification methods. Pei *et al.* [103] proposed a generic framework explicitly models an attacker to evaluate the security and robustness of ML systems. Wicker *et al.* [91] proposed a feature-guided "black-box" method to test the resilience of an image classifier against the adversarial example. The method requires neither knowledge of the network nor extensive sampling of the network to train a new one.

c. Robustness Guarantee

Several approaches to improve the robustness guarantee of DNN models have been proposed. In terms of formal guarantee of the classifier robustness, Hein and Andriushchenko [108] proposed the Cross-Lipschitz regularization functional approach for the first time in kernel methods to improve the robustness of the classifier without losing its prediction performance. Gopinath *et al.* [109] proposed DeepSafe, a data-guided approach that automatically identifies safety regions of input space to make the network robust to adversarial perturbations. Gehr *et al.* [102] proposed AI², the first sound and scalable analyzer for DNNs that can automatically prove its robustness. Hendrycks and Dietterich [105] first introduced a comprehensive corruption and perturbation robustness benchmark. Hein *et al.* [106] proposed a robust optimization technique that enforces low confidence prediction away from the training data, which can significantly reduce the confidence of noise images. Croce *et al.* [107] proposed a regularization scheme to improve the robustness of the ReLU network classifier by maximizing the linear region and the distance from the decision boundary.

d. Other Countermeasures

Wang *et al.* [104] proposed a detection algorithm that integrates mutation and hypothesis testing to detect adversarial examples at runtime. The algorithm hypothesis in most cases the adversarial examples in the DNN model are more sensitive than normal samples. In terms of structure inspection, Papernot and McDaniel [110] proposed the DkNN algorithm to inspect the internal DNN during testing, combining the k-nearest neighbor algorithm and the data representation of DNN learning at each level to provide confidence, interpretability, and robustness. Yuan *et al.* [111] firstly reviewed the latest research findings of DNN adversarial examples from the perspective of deep learning and proposed a taxonomy of adversarial generation methods. The study argued that all defences are only effective against local attacks and some strong, unseen attacks are often ineffective, thus there is an urgent need to build new defences, especially for safety-critical systems.

4.3.2.2 Rule Extraction for DNN

Rule extraction has the potential to extend ANN into safety-critical applications. Hailesilassie [112] comprehensively reviewed various rule extraction algorithms in neural networks and analyzed their applicability in DNN and argued that research on DNN rule extraction algorithms is limited and needs attention. In addition, pedagogical approach seems to be promising as it is not affected by the number of hidden layers and does not depend on the architecture of the algorithm.

4.4 Safe Learning

Reinforcement learning algorithms interact with the environment through trial-and-error, which makes them a very powerful paradigm for learning optimal policies. However, agents often explore all possible actions to find the optimal policy, sometimes even choosing actions at random. From a safety point of view, this can have serious consequences and that could be potentially harmful to real-world safety-critical systems [119]. At present, research on safe reinforcement learning is mainly aimed at this problem, mainly based on control theory, model predictive control (MPC), Hamilton-Jacobi-Isaacs (HJI) reachability analysis and the Markov decision process.

4.4.1 Control Theory

In order to avoid safety problems, some studies on proposing algorithms for nonlinear, closed-loop systems considering their ability to recover safely from exploratory actions based on control theory. Region of attraction (ROA) is an important property of nonlinear systems, a typical method to quantify ROA is to use the level set of Lyapunov function. Gaussian processes (GPs) are used to learn the dynamics of nonlinear systems from data to obtain high probability safety guarantees. Berkenkamp *et al.* [115] proposed a reinforcement learning algorithm based on the initial approximation model and the corresponding Lyapunov function, which learns the ROA from experiments of real systems without leaving the true ROA. As a result, it does not run the risk of safety-critical failures. Berkenkamp *et al.* [114] proposed another learning algorithm that provides safety optimization policies in continuous state-action spaces to achieve high-probability safety guarantees. Richards *et al.* [113] argued that safe learning should be guaranteed by verifying the safety certificate of a state prior to exploration. Based on this, the authors constructed a neural network Lyapunov function and further proposed a training algorithm adapted to the shape of the maximum safe region in the state space.

4.4.2 Model Predictive Control

MPC is a statistical modelling technique for estimating system uncertainty. Koller *et al.* [116] proposed a safety learning-based MPC framework to provide a high probability of safety guarantees throughout the learning process. Kahn *et al.* [120] proposed PLATO, which is a continuous, reset-free policy search algorithm that uses an adaptive training method to modify the MPC behavior to gradually match the learned policies and generate training samples.

4.4.3 Reachability-based Safe Learning and Assured Reinforcement Learning

Akametalu *et al.* [117] proposed a learning algorithm using GPs to learn system disturbance model and adopted a novel control strategy. Based on HJI reachability analysis, the algorithm determines and maintains a safe region in the state space by providing a control policy. For assured reinforcement learning, Mason *et al.* [119] proposed to modelling the uncertain environment as a high-level and abstract Markov decision process (AMDP), enabling an autonomous agent to solve decision making problems under constraints.

4.4.4 Verification Challenges

Wesel and Goodloe [118] discussed the challenges in the verification of ML algorithms and used a specific example to verify reinforcement learning algorithm. The author argued that there is no one-size-fit-all solution, different domains and algorithms allow different verification methods.

V. Discussion of Future Directions

Based on the above findings, in this section, some potential future research directions are discussed, so as to promote the application and development of AI in safety-critical systems.

5.1 Holistic Perspective

Managing safety problems require a holistic view and concerning knowledge involving different disciplines [137]. The application of AI in the safety-critical domain is a multi-disciplinary field, which combines the AI community, safety-critical domain, computer science and software engineering. The future development of AI in safety-critical systems needs to be looked at from a holistic perspective. Thinking and solving the problem from a holistic perspective across different disciplines may lead to new ideas. In addition, survey in [121] discusses the ML assurance in safety-critical systems from the perspective of the entire ML lifecycle, covering from data management to model learning, model verification and model deployment. Furthermore, the safety lifecycle view was adopted in the ANN model development of safety-critical systems [68]. As a result, the future development of AI in safety-critical systems should also be taken with the holistic perspective of the whole lifecycle, such as AI lifecycle, safety lifecycle, system lifecycle, software lifecycle and others. This is particularly important when building an AI model for a safety-critical system.

5.2 Verification & Validation

In the literature, V&V for different AI techniques has been studied. For interpretable methods, formal verification [33][34][35] has been studied for the decision tree and random forest and there are only two papers on V&V for BN model [64]. In terms of the ANN and DNN, V&V has been made more effort. With regard to safe learning, only one technical report addressed the verification challenge [118] and no in-depth research was conducted. More research is needed on the V&V of AI application in safety-critical systems, especially for interpretable methods, because the structure is more intuitive and simpler. For neural networks, V&V of the DNN has been studied more than other AI techniques, but the current verification is only applicable to a limited class. However, studies have shown that most verifications are not robust and scalable [100], that is, such neural networks are vulnerable to future attacks, and current verification methods are limited by the network size. Some research on scalable V&V [135][136] may be a future direction of research. In addition, the existing DNN verification mainly focuses on adversarial examples, whereas future research can extend to a broader perspective.

5.3 Formal Methods

Formal methods have been widely discussed in the literature, including formal verification, safety analysis [98] and formal guarantee [108]. Formal methods can guarantee fault-free development of the system [138], which is a way to increase confidence and the unification and harmonization of engineering practices involved in building software-based safety-critical systems [139][141]. The use of formal methods in safety-critical systems will result in accurate, consist, and correctness of the proposed system [140]. Therefore, the future development of AI in safety-critical systems requires more formal methods plus the application of formal verification techniques. In addition, formal methods can be combined with the field of AI, and this holds a great potential for the use of AI in safety-critical systems.

5.4 Awareness

The usage of the term “safety-critical” in academia was not very active. The term often appeared in articles on autonomous vehicles or adversarial examples. Articles regarding the application of BN to safety-critical systems on process plants, process systems or subsea BOP systems are all safety-critical, the term was not used. This is particularly evident in articles prior to 2015. Raising awareness of its usage in the literature and establishing safety policies and objectives at the organizational level [142] can in term avoid the phenomenon of article fragmentation in the literature and further promote the development of the field.

VI. CONCLUSIONS

Safety-critical systems are receiving increasingly attention, one trend with ample potential is the adoption of AI. This paper conducted a systematic review of 92 papers and analyzed the application of AI in safety-critical systems, which lays a foundation for future research in this domain.

The methodology use in this paper contains four stages: planning, selection, extraction and execution. Based on the thematic analysis, three themes were identified: interpretable method, explain model behaviour and safe learning. Interpretable method was further divided into two subcategories: decision tree & random forest and BN. Research on the former has just started and there have been only 3 papers on formal verification. There were total 35 papers on BN applications in system dependability, reliability, availability and safety analysis, V&V, system design and abnormal situation management. Explaining model behaviour focused on ANNs and DNNs. Given that the DNN literature is based on the problem of adversarial examples, as an intense research topic it has been separate for discussion in this paper. There were total 17 papers on ANNs, covering V&V and model development. Research of V&V was mainly on the "black-box" problem, one solution is obtaining explanation capability through rule extraction. ANN model development was mainly focused on building ANN and fuzzy logic hybrid model using rule extraction to gain "white-box" analysis and transparency. For DNNs, there were a total of 28 papers on adversarial examples, from the aspects of testing, verification, robustness guarantee, structure inspection and detection algorithm. Among them, formal verification was the focus of the research. Safe learning is about using reinforcement learning, which has rarely been applied to safety-critical applications due to safety issues. There were 8 papers in this category and concerning safe learning algorithms based on control theory, model predictive control (MPC), Hamilton-Jacobi-Isaacs (HJI) reachability analysis, and the Markov decision process. In addition, 1 survey paper regarding ML assurance of safety-critical systems has been discussed.

Finally, on the basis of the literature review and analysis, the authors proposed four research directions for future development of AI in safety-critical systems. Future research needs to be from a holistic multi-disciplinary and lifecycle view and more research on V&V is needed, especially scalable V&V. Moreover, formal methods have a great potential in the application of AI in safety-critical systems, and the development of AI formal methods is promising. There is also a need to raise awareness of using the term "safety-critical", which will contribute to promote the domain development.

REFERENCES

[1] Pannu, A. (2015). Artificial Intelligence and its Application in Different Areas. *International Journal of Engineering and Innovative Technology (IJEIT)*, 4(10), 79–84.

[2] Hirschberg, J. and Manning, C.D. (2015). Advances in natural language processing. *Science*, 349(6245), 261–266.

[3] Voulodimos, A., Doulamis, N., Doulamis, A. and Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018, 1–13.

[4] Jun Lee, S. and Siau, K. (2001). A review of data mining techniques. *Industrial Management & Data Systems*, 101(1), 41–46.

[5] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mane,D. (2016). Concrete problems in AI safety. *arXiv preprint*, arXiv:1606.06565.

[6] Knight, J.C. (2002). Safety-Critical Systems: Challenges and Directions. *Proceedings of ICSE'2002: 24th International Conference on Software Engineering*, ACM Press, 547-550.

[7] Bowen, J. and Stavridou, V. (1993). Safety-critical systems, formal methods and standards. *Software Engineering Journal*, *8*(4), 189.

[8] Kohli, P. and Chadha, A. (2020). Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving car crash. *Lecture Notes in Networks and Systems*, Vol 69, 261-279.

[9] Tian, Y., Pei, K., Jana, S. and Ray, B. (2018). DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars. *International Conference on Software Engineering (ICSE). ACM*.

[10] Tranfield, D., Denyer, D. and Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, 14(3), 207–222.

[11] Winfield, A. F.T. and Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180085.

[12] Jobin, A., Ienca, M. and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

[13] Russell, S., Dewey, D. and Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, *36*(4), 105–114.

[14] Seshia, S.A., Sadigh, D. and Sastry, S.S. (2016). Towards verified artificial intelligence. *arXiv preprint*, arXiv:1606.08514.

[15] Baldassarre, M. T., Caivano, D., Kitchenham, B. and Visaggio, G. (2007). Systematic Review of Statistical Process Control: An Experience Report. *Proceedings of 11th International Conference Evaluation and Assessment in Software Engineering, (EASE 07), British Computer Society*, 1–9.

[16] Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... Williams, M.D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57(101994), 1-47.

[17] Nanda, M. and Jayanthi, J. (2013). An Effective Verification and Validation Strategy for Safety-Critical Embedded Systems. *International Journal of Software Engineering & Applications*, *4*(2), 123–142.

[18] Zednik, C. (2019). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 1-24.

[19] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B. and Swami, A. (2017). Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp.506–519.

[20] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

[21] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M. and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *Proceedings of IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, 80–89.

[22] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

[23] Doshi-Velez, F. and Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints Ml*, 1–12.

[24] Goodman, B. and Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine, 38*(3), 50–57.

[25] Ward, F.R. and Habli, I. (2020). An assurance case pattern for the interpretability of machine learning in safety-critical systems. *SAFECOMP 2020 Workshops*, Springer Int. Publishing, 395–407.

[26] IEC61508: Functional safety of electrical/electronic/programmable electronic safety-related systems. *International Electrotechnical Commission* (Ed 2, April 2010)

[27] Guiochet, J., Do Hoang, Q.A., Kaaniche, M. and Powell, D. (2012). Applying existing standards to a medical rehabilitation robot: Limits and challenges. Workshop FW5: Safety in Human-Robot Coexistence & Interaction: How can Standardization and Research benefit from each other?, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2012)*.

[28] Floridi, L. (2020). AI and Its New Winter: from Myths to Realities. *Philosophy & Technology*, 33, 1–3.

[29] Cummings, M.L. (2021). Rethinking the maturity of artificial intelligence in safety-critical settings. *AI Magazine*, 1-13.

[30] Yasnitsky, L.N. (2020). Whether Be New "Winter" of Artificial Intelligence?. Antipova T. (eds) Integrated Science in Digital Age. ICIS 2019. *Lecture Notes in Networks and Systems*, vol 78. Springer, Cham.

[31] Mehta, N. and Devarakonda, M.V. (2018). Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey? *Journal of Allergy and Clinical Immunology, 141*(6), 2019–2021.e1.

[32] Bostrom, N. and Yudkowsky, E. (2014).The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence* (eds. Frankish, K. & Ramsey, W. M.), Cambridge University Press, 316–334.

[33] Törnblom, J. and Nadjm-Tehrani, S. (2018). Formal verification of random forests in safety-critical applications. *Int. Workshop on Formal Techniques for Safety-Critical Systems*. Springer, 55–71.

[34] Törnblom, J. and Nadjm-Tehrani, S. (2019). An abstraction-refinement approach to formal verification of tree ensembles. *Proceedings of 2nd Int. Workshop on Artificial Intelligence Safety Engineering*, held with SAFECOMP. Springer.

[35] Törnblom, J. and Nadjm-Tehrani, S. (2019). Formal verification of input-output mappings of tree ensembles. *arXiv preprint*, arXiv: 1905.04194.

[36] Torres-Toledano, J. and Sucar, L. (1998). Bayesian networks for reliability analysisof complex systems. *Proceedings of the sixth Ibero-American conference on AI (IBERAMIA 98)*, no. 1484 in lecture notes in artificial intelligence, Berlin, Germany: Springer, 195–206.

[37] Langseth, H. and Portinale, L. (2007). Bayesian networks in reliability. *Reliability Engineering & System Safety*, 92(1), 92–108.

[38] Simon, C., Weber, P. and Evsukoff, A. (2008). Bayesian networks inference algorithm to implement Dempster Shafer theory in reliability analysis. *Reliability Engineering & System Safety*, 93(7), 950–963.

[39] Simon, C., Weber, P. and Levrat, E. (2007). Bayesian Networks and Evidence Theory to Model Complex Systems Reliability. *Journal of Computers*, 2(1).

[40] Montani, S., Portinale, L., Bobbio, A. and Codetta-Raiteri, D. (2008). Radyban: A tool for reliability analysis of dynamic fault trees through conversion into dynamic Bayesian networks. *Reliability Engineering & System Safety*, 93(7), 922–932.

[41] Amrin, A., Zarikas, V. and Spitas, C. (2018). Reliability analysis and functional design using Bayesian networks generated automatically by an "Idea Algebra" framework. *Reliability Engineering & System Safety*, 180, 211–225.

[42] Weber, P. and Jouffe, L. (2003). Reliability modelling with dynamic bayesian networks. *IFAC Proceedings Volumes*, 36(5), 57–62.

[43] Weber, P. and Jouffe, L. (2006). Complex system reliability modelling with Dynamic Object Oriented Bayesian Networks (DOOBN). *Reliability Engineering & System Safety*, 91(2), 149–162.

[44] Boudali, H. and Dugan, J. (2005). A discrete-time Bayesian network reliability modeling and analysis framework. *Reliability Engineering & System Safety*, 87(3), 337–349.

[45] Doguc, O. and Ramirez-Marquez, J. E. (2009). A generic method for estimating system reliability using Bayesian networks. *Reliability Engineering & System Safety*, 94(2), 542–550.

[46] Cai, B., Liu, Y., Liu, Z., Tian, X., Dong, X. and Yu, S. (2012). Using Bayesian networks in reliability evaluation for subsea blowout preventer control system. *Reliability Engineering & System Safety*, 108, 32–41.

[47] Liu, Z. and Liu, Y. (2019). A Bayesian network based method for reliability analysis of subsea blowout preventer control system. *Journal of Loss Prevention in the Process Industries*, 59, 44–53.

[48] Yang, Z., Bonsall, S. and Wang, J. (2008). Fuzzy Rule-Based Bayesian Reasoning Approach for Prioritization of Failures in FMEA. *IEEE Transactions on Reliability*, 57(3), 517–528.

[49] Cai, B., Huang, L. and Xie, M. (2017). Bayesian networks in Fault Diagnosis. *IEEE Transctions on Industrial Informatics*, 13(5): 2227-40.

[50] Khakzad, N., Khan, F. and Amyotte, P. (2011). Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches. *Reliability Engineering & System Safety*, 96(8), 925–932.

[51] Khakzad, N., Khan, F. and Amyotte, P. (2013). Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. *Process Safety and Environmental Protection*, 91(1–2), 46–53.

[52] Stefaniak, B., Cholewiński, W. and Tarkowska, A. (2005). Algorithms of Artificial Neural Networks - Practical application in medical science. Polski Merkuriusz Lekarski,19,819-22.

[53] Gran, B. A. (2002). Use of Bayesian Belief Networks when combining disparate sources of information in the safety assessment of software-based systems. *International Journal of Systems Science*, 33(6), 529–542.

[54] Prabhakaran, R., Krishnaprasad, R., Nanda, M. and Jayanthi, J. (2016). System Safety Analysis for Critical System Applications Using Bayesian Networks. *Procedia Computer Science*, 93, 782–790.

[55] Neil, M., Littlewood, B. and Fenton, N. (1996). Applying Bayesian Belief Networks to Systems Dependability Assessment. *Proceedings of Safety Critical Systems Club Symp., Springer-Verlag, Leeds*.

[56] Fenton, N., Littlewood, B., Neil, M., Strigini, L., Sutcliffe, A. and Wright, D. (1998). Assessing dependability of safety critical systems using diverseevidence. *IEEE Proceedings of Software Engng 1998*, 145(1):35–9.

[57] Kang, C. W. and Golay, M. W. (1999). A Bayesian belief network-based advisory system for operational availability focused diagnosis of complex nuclear power systems. *Expert Systems with Applications*, 17(1), 21–32.

[58] Bobbio, A., Portinale, L., Minichino, M. and Ciancamerla, E. (2001). Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliability Engineering & System Safety*, 71(3), 249–260.

[59] Montani, S., Portinale, L. and Bobbio, A. (2005). Dynamic Bayesian networks formodeling advanced fault tree features in dependability analysis. *Proceedings of the sixteenth European conference on safetyand reliability. Leiden, The Netherlands: A.A. Balkema*, 1415–22.

[60] Amin, M. T., Khan, F. and Imtiaz, S. (2018). Dynamic availability assessment of safety critical systems using a dynamic Bayesian network. *Reliability Engineering & System Safety*, 178, 108–117.

[61] Bouissou, M., Martin, F. and Ourghanlian, A. (1999). Assessment of a safety criticalsystem including software: a Bayesian belief network for evidence sources. *Reliability and Maintainability Symposium (RAMS'99), Washington, January*.

[62] Naderpour, M., Lu, J. and Zhang, G. (2014). An intelligent situation awareness support system for safety-critical environments. *Decision Support Systems*, 59, 325–340.

[63] Naderpour, M., Lu, J. and Zhang, G. (2015). An abnormal situation modeling method to assist operators in safety-critical systems. *Reliability Engineering & System Safety*, 133, 33–47.

[64] Douthwaite, M. and Kelly, T. (2017). Establishing Verification and Validation Objectives for Safety-Critical Bayesian Networks. *International Symposium on Software Reliability Engineering (ISSRE) '17 - Special Session on AI (Workshop on Software Certification), IEEE*.

[65] Kannan, P. R. (2007). Bayesian networks: Application in safety instrumentation and risk reduction. *ISA Transactions*, 46(2), 255–259.

[66] Weber, P., Medina-Oliva, G., Simon, C. and Iung, B. (2012). Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, 25(4), 671–682.

[67] Kabir, S. and Papadopoulos, Y. (2019). Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. *Safety Science*, 115, 154–175.

[68] Kurd, Z. and Kelly, T. (2003). Safety Lifecycle for Developing Safety-critical Artificial Neural Networks. *22nd International Conference on Computer Safety, Reliability and Security (SAFECOMP'03), 23-26 September*.

[69] Kurd, Z., Kelly, T. and Austin, J. (2007). Developing artificial neural networks for safety critical systems. *Neural Computing and Applications*, 16(1), 11–19.

[70] Kelly, T. (2004). A systematic approach to safety case management. *Proceedings of Society of Automotive Engineers (SAE) World Congress, March*, SAE transactions.

[71] Rodvold, D. M. (1999). A software development process model for artificial neural networks in critical applications. Neural Networks, 1999. *IJCNN'99. International Joint Conference*, vol. 5. IEEE, 3317–3322.

[72] Weaver, R. A., McDermid, J. A. and Kelly, T. P. (2002). Software Safety Arguments: Towards a Systematic Categorisation of Evidence. *Proceedings of the 20th International System Safety Conference (ISSC 2002), System Safety Society, Denver, Colorado, USA*.

[73] Kurd, Z. and Kelly, T. (2003). Establishing safety criteria for artificial neural networks. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, 163-169.

[74] Ward, F. R. and Habli, I. (2020). An Assurance Case Pattern for the Interpretability of Machine Learning in Safety-Critical Systems. *Third International Workshop on Artificial Intelligence Safety Engineering*.

[75] Wen, W., Callahan, J. and Napolitano, M. (1996). Towards developing verifiable neural network controllers. *Proceedings of the Workshop on AI for Aeronautics and Space*.

[76] Kurd, Z. and Kelly, T. P. (2007). Using fuzzy self-organising maps for safety critical systems. *Reliability Engineering & System Safety*, 92(11), 1563–1583.

[77] Kurd, Z. and Kelly, T. P. (2005). Using safety critical artificial neural networks ingas turbine aero-engine control. *Presented at 24th internationalconference on computer safety, reliability and security (SAFE-COMP'05), 28–30 September, Fredrikstad, Norway*.

[78] Bedford, D., Morgan, G. and Austin, J. (1996). Requirements for a standard certifying the use of artificial neural networks in safety critical applications. *Proceedings of the international conference on artificial neural networks*.

[79] Zhang, J. and Li, C. (2019). Adversarial Examples: Opportunities and Challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 1–16.

[80] Taylor, B., Darrah, M. and Moats, C. (2003). Verification and validation of neural networks: a sampling of research in progress. *Proceedings of Society of Photo-optical Instrumentation Engineers*, 8-16.

[81] Andrews, R., Diederich, J. and Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6), 373–389.

[82] Taylor, B. J. and Darrah, M. A. (2005). Rule extraction as a formal method for the verification and validation of neural networks. *IEEE International Joint Conference on Neural Networks* 5, 2915–2920.

[83] Hull, J., Ward, D. and Zakrzewski, R. R. (2002). Verification and Validation of Neural Networks for SafetyCritical Applications. *Proceedings of American Control Conference, Anchorage, AK*.

[84] Darrah, M., Taylor, B., Webb, M. and Livingston, R. (2005). A Geometric Rule Extraction Approach used for Verification and Validation of a Safety Critical Application. *2005 Florida Artificial Intelligence Research Society Conference, Clear Water Beach, FL, May 16-18*.

[85] Darrah, M., Taylor, B. and Skias, S. (2004). Rule Extraction from Dynamic Cell Structure Neural Networks Used in a Safety Critical Application. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, 629-634.

[86] Ma, L., Juefei-Xu, F., Sun, J., Chen, C., Su, T., Zhang, F., Xue, M., Li, B., Li, L., Liu, Y. and others. (2018). DeepGauge: Comprehensive and Multi-Granularity Testing Criteria for Gauging the Robustness of Deep Learning Systems. *arXiv preprint*, arXiv:1803.07519.

[87] Ma, L., Zhang, F., Sun, J., Xue, M., Li, B., Juefei-Xu, F., Xie, C., Li, L., Liu, Y., Zhao, J. and Wang, Y. (2018). DeepMutation: Mutation Testing of Deep Learning Systems. *International Symposium on Software Reliability Engineering (ISSRE)*.

[88] Pei, K., Cao, Y., Yang, J. and Jana,S. (2017). Deepxplore: Automated whitebox testing of deep learning systems. *Proceedings of the ACM SIGOPS 26th symposium on Operating systems principles*.

[89] Sun, Y., Huang, X and Kroening, D. (2018). Testing Deep Neural Networks. *arXiv preprint*, arXiv:1803.04792.

[90] Sun, Y., Wu, M., Ruan, W., Huang, X., Kwiatkowska, M. and Kroening, D. (2018). Concolic testing for deep neural networks. *Automated Software Engineering (ASE), 33rd IEEE/ACM International Conference*.

[91] Wicker, M., Huang, X. and Kwiatkowska, M. (2018). Feature-guided black-box safety testing of deep neural networks. *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*.

[92] Zhang, J. M., Harman, M., Ma, L. and Liu, Y. (2020). Machine Learning Testing: Survey, Landscapes and Horizons. *IEEE Transactions on Software Engineering*, 1-37.

[93] Huang, X., Kwiatkowska, M., Wang, S. and Wu, M. (2017). Safety verification of deep neural networks. *Computer Aided Verification: 29th International Conference (CAV)*, 3–29.

[94] Katz, G., Barrett, C., Dill, D. L., Julian, K. and Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. *Computer Aided Verification: 29th International Conference (CAV)*, 97–117.

[95] Pulina, L. and Tacchella, A. (2010). An Abstraction-Refinement Approach to Verification of Artificial Neural Networks. *Computer Aided Verification*, Springer, Berlin, Heidelberg, 243-257.

[96] Cheng, C-H, Nuhrenberg, G. and Ruess, H. (2017). Maximum resilience of artificial neural networks. *International Symposium on Automated Technology for Verification and Analysis*, Springer, 251–268.

[97] Bunel, R., Turkaslan, I., Torr, P.H.S., Kohli, P. and Kumar, M.P. (2018). A Unified View of Piecewise Linear Neural Network Verification. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*.

[98] Wang, S., Pei, K., Whitehouse, J., Yang, J. and Jana, S. (2018). Efficient formal safety analysis of neural networks. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*.

[99] Xiang, W., Tran, H-D. and Johnson, T-T. (2017). Reachable set computation and safety verification for neural networks with ReLU activations. *arXiv preprint*, arXiv: 1712.08163.

[100] Carlini, G., Katz, G., Barrett, C. and Dill, D. (2017). Provably minimally-distorted adversarial examples. arXiv *preprint*, arXiv: 1711.00851.

[101] Xiang, W., Tran, H. D. and Johnson, T. T. (2018). Output Reachable Set Estimation and Verification for Multilayer Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11), 5777–5783.

[102] Gehr, T., Mirman, M., Draschler-Cohen, D., Tsankov, P., Chaudhuri, S. and Vechev, M. (2018). AI2 : Safety and Robustness Certification of Neural Networks with Abstract Interpretation. *IEEE Symposium on Security and Privacy*, 39.

[103] Pei, K., Cao, Y., Yang, J. and Jana, S. (2017). Towards practicacl verification of machine learning: The case of computer vision systems. *arXiv preprint*, arXiv: 1712.01785.

[104] Wang, J., Dong, G. Sun, J., Wang, X. and Zhang, P. (2019) Adversarial sample detection for deep neural network through model mutation testing. *Proceedings of the 41st International Conference on Software Engineering (ICSE), IEEE*, 1245–1256.

[105] Hendrycks, D. and Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *International Conference on Learning Representations (ICLR)*.

[106] Hein, M., Andriushchenko, M. and Bitterwolf, J. (2019). Why ReLU networks yield highconfidence predictions far away from the training data and how to mitigate the problem. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 41–50.

[107] Croce, F., Andriushchenko, M. and Hein, M. (2019). Provable robustness of relu networks via maximization of linear regions. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

[108] Hein, M. and Andriushchenko, M. (2017). Formal guarantees on the robustness of a classifier against adversarial manipulation. *arXiv preprint*, arXiv: 1705.08475.

[109] Gopinath, D., Katz, G., Păsăreanu, C. S. and Barrett, C. (2017). Deepsafe: A data-driven approach for checking adversarial robustness in neural networks. *arXiv preprint*, arXivL: 1710.00486.

[110] Papernot, N. and McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint*, arXiv:1803.04765.

[111] Yuan, X., He, P., Zhu, Q. and Li, X. (2019). Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9), 2805–2824.

[112] Hailesilassie, T. (2016). Rule extraction algorithm for deep neural networks: a review. *International Journal of Computer Science and Information Security*, 14(7), 376-381.

[113] Richards, S. M., Berkenkamp, F. and Krause, A. (2018). The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of Proceedings of Machine Learning Research, 466–476.

[114] Berkenkamp, F., Turchetta, M., Schoellig, A. P. and Krause, A. (2017). Safe Model-based Reinforcement Learning with Stability Guarantees. *Advances in Neural Information Processing Systems*.

[115] Berkenkamp, F., Moriconi, R., Schoellig, A. P. and Krause, A. (2016). Safe learning of regions of attraction for uncertain, nonlinear systems with Gaussian processes. *Proceedings of the IEEE Conference on Decision and Control*.

[116] Koller, T., Berkenkamp, F., Turchetta, M. and Krause, A. (2018). Learning-based model predictive control for safe exploration and reinforcement learning. *arXiv preprint*, arXiv:1803.08287.

[117] Akametalu, A. K., Fisac, J. F., Gillula, J. H., Kaynama, S., Zeilinger, M. N. and Tomlin, C. J. (2014). Reachability-based safe learning with Gaussian processes. *53rd IEEE Conference on Decision and Control*.

[118] Wesel, P. and Goodloe, A. E. (2017). Challenges in the Verification of Reinforcement Learning Algorithms. *NASA Technical Report*, NASA/TM–2017–219628.

[119] Mason, G, Calinescu, R., Kudenko, D. and Banks, A. (2017). Assured reinforcement learning with formally verified abstract policies. *9th International Conference on Agents and Artificial Intelligence*.

[120] Kahn, G., Zhang, T., Levine, S. and Abbeel, P. (2017). Plato: Policy learning using adaptive trajectory optimization. *International Conference on Robotics and Automation*.

[121] Ashmore, R., Calinescu, R. and Paterson, C. (2019). Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *arXiv preprint*, arXiv: 1905.04223.

[122] Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13), 2047–2054.

[123] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. and Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6), 275–285.

[124] Okoli, C. and Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Sprouts: Working Papers on Information Systems*, 10(26).

[125] Biau, G. and Scornet, E. (2016). Rejoinder on: A random forest guided tour. *TEST*, 25(2), 264–268.

[126] Grimm, T., Lettnin, D. and Hübner, M. (2018). A Survey on Formal Verification Techniques for Safety-Critical Systems-on-Chip. *Electronics*, 7(6), 81.

[127] Wiels, V., Delmas, R., Doose, D., Garoche, P.-L., Cazin, J. and Durrieu, G. (2012). Formal Verification of Critical Aerospace Software. *AerospaceLab Journal*, Issue 4.

[128] MacKenzie, D. & Pottinger, G. (1997). Mathematics, technology, and trust: formal verification, computer security, and the U.S. military. *IEEE Annals of the History of Computing*, 19(3), 41–59.

[129] Johnson, T. L. (2004). Improving Automation Software Dependability: A Role for Formal Methods? *IFAC Proceedings Volumes*, 37(4), 153–164.

[130] Pasman, H. and Rogers, W. (2013). Bayesian networks make LOPA more effective, QRA more transparent and flexible, and thus safety more definable! *Journal of Loss Prevention in the Process Industries*, 26(3), 434–442.

[131] Yang, Z., Bonsall, S. and Wang, J. (2009). Use of hybrid multiple uncertain attribute decision making techniques in safety management. *Expert Systems with Applications*, 36(2), 1569–1586.

[132] Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W. and Goyal, H. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine*, 6(11), 216.

[133] Nusser, S. (2009). Robust learning in safety-related domains: machine learning methods for solving safetyrelated application problems. *Doctoral dissertation, Otto-von-Guericke-Universita ̈t Magdeburg*. Available at: https://pdfs.semanticscholar.org/48c2/e5641101a4e5250ad903828c02025d269a1a.pdf.

[134] Guo, F., Zhao, Q., Li, X., Kuang, X., Zhang, J., Han, Y. and Tan, Y. A. (2019). Detecting adversarial examples via prediction difference for deep neural networks. *Information Sciences*, 501, 182–192.

[135] Krishnamurthy, D., Robert, S., Sven, G., Timothy, M. and Pushmeet, K. (2018). A dual approach to scalable verification of deep networks. *Conference on Uncertainty in Artificial Intelligence*.

[136] Kuper, L., Katz, G. Gottschlich, J. Julian, K. Barrett, C. and Kochenderfer, M. (2018) Toward Scalable Verification for Safety-Critical Deep Networks. *Association for Uncertainty in Artificial Intelligence*.

[137] Aven, T. and Ylönen, M. (2018). A risk interpretation of sociotechnical safety perspectives. *Reliability Engineering & System Safety*, 175, 13–18.

[138] Idani, A., Ledru, Y., Wakrime, A-A., Ayed, R-B. and Dutilleul, S-C. (2019). Incremental development of a safety critical system combining formal methods and dsmls: application to a railway system. 24th international conference on formal methods for industrial critical systems. *LNCS*, vol. 11687, Springer, 93–109.

[139] Bowen, J. and Stavridou, V. (1993). Safety-critical systems, formal methods and standards. *Software Engineering Journal*, 8(4), 189.

[140] Singh, M., Sharma, A. K. and Saxena, R. (2015). Why Formal Methods Are Considered for Safety Critical Systems? *Journal of Software Engineering and Applications*, 08(10), 531–538.

[141] Wang, J. (2000). Analysis of safety-critical software elements in offshore safety studies. *Disaster Prevention and Management*, 9(4), 271–282.

[142] Law, W., Chan, A. and Pun, K. (2006). Prioritising the safety management elements. *Industrial Management & Data Systems*, 106(6), 778–792.

[143] Luo, S., Liu, H. and Qi, E. (2019). Big data analytics – enabled cyber-physical system: model and applications. *Industrial Management & Data Systems*, 119(5), 1072–1088.

[144] Raut, R., Narwane, V., Kumar Mangla, S., Yadav, V. S., Narkhede, B. E. and Luthra, S. (2021). Unlocking causal relations of barriers to big data analytics in manufacturing firms. *Industrial Management & Data Systems*, ahead-of-print.

[145] Maguire, M and Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education*, 9(3).

[146] Cai, B., Liu, Y. and Xie, M. (2017). A Dynamic-Bayesian-Network-Based Fault Diagnosis Methodology Considering Transient and Intermittent Faults. *IEEE Transactions on Automation Science and Engineering*, 14(1), 276–285.

[147] Wang, C., Liu, Y., Hou, W., Wang, G. and Zheng, Y. (2020). Reliability and availability modeling of Subsea Xmas tree system using Dynamic Bayesian network with different maintenance methods. *Journal of Loss Prevention in the Process Industries*, 64, 104066.

[148] Chiremsel, Z., Nait Said, R. and Chiremsel, R. (2016). Probabilistic Fault Diagnosis of Safety Instrumented Systems based on Fault Tree Analysis and Bayesian Network. *Journal of Failure Analysis and Prevention*, 16(5), 747–760.

[149] Schietekat, S., De Waal, A. and Gopaul, K.G. (2016). Validation & verification of a Bayesian network model for aircraft vulnerability. *12th INCOSE SA Systems Engineering Conference*, 12-14 September 2016, CSIR International Convention Centre, Pretoria, South Africa.

[150] Schwendicke, F., Samek, W. and Krois, J. (2020). Artificial Intelligence in Dentistry: Chances and Challenges. *Journal of Dental Research*, 99(7), 769–774.

[151] Dobrev, D. (2005). A Definition of Artificial Intelligence. *Mathematica Balkanica*, New Series, Vol. 19, Fasc. 1-2, 67-74.

[152] Dobrev, D. (2005). Formal definition of artificial intelligence. *International Journal of Information Theories and Applications*, 12, 277–285.

[153] McCarthy, J. (2004). What is artificial intelligence? Available at: http://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf

[154] Yampolskiy, R. V. and Spellchecker, M. S. (2016). Artificial intelligence safety and cybersecurity: a timeline of ai failures. *arXiv preprint*, arXiv:1610.07997.

[155] Johnson, D. M. (1996). A review of fault management techniques used in safety-critical avionic systems. *Progress in Aerospace Sciences*, 32(5), 415–431.

[156] Kumar, P., Singh, L. K. and Kumar, C. (2020). Performance evaluation of safety-critical systems of nuclear power plant systems. *Nuclear Engineering and Technology*, 52(3), 560–567.

[157] Weissnegger, R., Schuss, M., Kreiner, C., Pistauer, M., Römer, K. and Steger, C. (2016). Simulation-based Verification of Automotive Safety-critical Systems Based on EAST-ADL. *Procedia Computer Science*, 83, 245–252.

[158] Fei, B., Ng, W. S., Chauhan, S. and Kwoh, C. K. (2001). The safety issues of medical robotics. *Reliability Engineering & System Safety*, 73(2), 183–192.

[159] Kriaa, S., Pietre-Cambacedes, L., Bouissou, M. and Halgand, Y. (2015). A survey of approaches combining safety and security for industrial control systems. *Reliability Engineering & System Safety*, 139, 156–178.

[160] Carson, J. S. (2002). Model verification and validation. *Proceedings of the Winter Simulation Conference*, vol.1, 52-58.

[161] Kim, Y.J. and Kim, M. (2012). Hybrid Statistical Model Checking Technique for Reliable Safety Critical Systems. *Proceedings of 23rd IEEE International Symposium on Software Reliability Engineering. IEEE*. 51–60.

[162] Fan, D. D., Nguyen, J., Thakker, R., Alatur, N., Agha-mohammadi, A.-a. and Theodorou, E. A. (2019). Bayesian learning-based adaptive control for safety critical systems. *arXiv preprint*, arXiv:1910.02325, 2019.

[163] Hsiung, P., Chen, Y and Lin, Y. (2007). Model Checking Safety-Critical Systems Using Safecharts. *IEEE Transactions on Computers*, 56(5), 692-705.

[164] Baldoni, R., Montanari, L. and Rizzuto, M. (2015). On-line failure prediction in safety-critical systems. *Future Generation Computer Systems*, 45, 123–132.

[165] Vassev E. (2016). Safe Artificial Intelligence and Formal Methods. Margaria T., Steffen B. (eds) Leveraging Applications of Formal Methods, Verification and Validation: Foundational Techniques. *Springer, Cham*. Lecture Notes in Computer Science, 9952.

[166] Righetti, L., Madhavan, R. and Chatila, R. (2019). Unintended Consequences of Biased Robotic and Artificial Intelligence Systems [Ethical, Legal, and Societal Issues]. IEEE Robotics & Automation Magazine, 26(3), 11–13.

[167] Yapo, A. and Weiss, J. (2018). Ethical implications of bias in machine learning. *Proceedings of the 51st Hawaii International Conference on System Sciences Big Island*. Hawaii, USA, 5365–5372.

[168] Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, 37.