IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 July 2024, accepted 5 August 2024, date of publication 8 August 2024, date of current version 24 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3440647

SURVEY

A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions

WISSAM SALHAB, DARINE AMEYED, FEHMI JAAFAR, AND HAMID MCHEICK, (Senior Member, IEEE)
Department of Computer Science and Mathematics, University of Québec at Chicoutimi, Chicoutimi, QC G7H 2B1, Canada
Corresponding author: Wissam Salhab (wissam.salhab1@uqac.ca)

ABSTRACT **Artificial intelligence (AI)** is revolutionizing many aspects of our lives, except it raises fundamental safety and ethical issues. In this survey paper, we review the current state of research on safe and trustworthy AI. This work provides a structured and systematic overview of AI safety. In which, we emphasize the significance of designing AI systems with safety focus, encompassing elements from data management, model development, and deployment. We underscore the need for AI systems to align with human values and operate within mounted ethical frameworks. In addition, we notice the need for a complete safety framework that courses the development and implementation of AI systems, ensuring they do not inadvertently cause damage to humans. Our results show that AI safety is associated with model learning techniques, verification and validation methods, failure modes, and managing AI autonomy. As discussed in the literature, the main concerns include explainability, interpretability, robustness, reliability, fairness, bias, and adversarial attacks.

INDEX TERMS AI safety, robustness, trustworthy AI, fairness and bias, explainability and interpretability.

I. INTRODUCTION

AI is a rapidly evolving field that aims to make computers or machines capable of carrying out tasks that typically need human intelligence [1]. These include the abilities to learn, think, solve problems, perceive, and language processing [2]. AI is frequently referred to as machine intelligence to distinguish it from human intelligence [3], [4], computer science and cognitive science came together to form the field [5].

With the advancement of AI technology, researchers claim that quality and safety are neglected and misunderstood in a rush to improve system performance by some metrics often linked to financial or market-related factors [6].

Meanwhile, there is a wide variety of perspectives on AI and many more on AI safety. Safety engineering professionals worry that traditional principles may not suffice for systems that learn and adapt during operation [7]. They also fear these systems can't be validated against a precise specification in which safety and control are difficult to guarantee [8].

Ensuring the safety of AI systems is essential, primarily when utilized in safety-critical systems [9], [10]. Creating trustworthy, safe, and reliable AI systems is necessary to avoid human harm or discrimination against individuals or groups [11], [12]. Trustworthiness is crucial for safety-critical applications because it ensures that predictions made by a model can be trusted in scenarios where every choice has consequences [13], [14], [15], [16], [17].

By prioritizing the system's safety and ensuring that these systems are transparent, predictable, and controllable by design, researchers can increase the trust and confidence in AI systems [18].

AI models acquire knowledge through various techniques that may involve processing and analyzing large volumes of data to identify patterns, anticipate outcomes, and make

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

predictions. Various AI learning processes, such as reinforcement learning, supervised learning, and unsupervised learning, are applied.

Machine learning (ML) is a subfield of AI constantly evolving and trying to mimic human intelligence through environmental learning [19]. ML is becoming common in cyber-physical systems [20]; these systems frequently use AI and ML techniques to function well in open contexts [21]. Because of their complexity, ensuring the safety and robustness of these systems is a significant problem. Therefore, the hazard may be presented in each of the life-cycle components: from data management, model development and analysis, and finally model deployment and monitoring [22], [23].

Many studies have discussed the usage of ML in safety-critical systems characterized by high levels of autonomy and safety. They may cause risks to human lives if misused [24]. These systems are deemed "safety-critical" or "high assurance" because they pose a risk of serious injury or financial loss in the event of failure [25], [26]. For example, in autonomous vehicles or healthcare [19], as well as solving problems in safety-critical systems using heterogeneous ensembles [27].

Therefore, the assurance of ML-based systems is one of the problems in safety-critical systems, where the goal is to provide a comprehensive and automated suite of processes for employing ML. This includes offering advice on how to design appropriate and safe testing criteria to gauge confidence in testing ML-based systems, such as how to generate tests efficiently, how to assess the quality and the size of test data, and how to ensure the robustness of ML models [23]

On the other hand, Reinforcement Learning (RL) is a branch of machine learning that studies optimizing the cumulative reward of intelligent agents acting in a given environment. [28]. In contrast to normal supervised learning, RL focuses on presenting, analyzing, and manipulating current knowledge that does not explicitly correct sub-optimal behaviors or present proper input/output pairs [29]. RL agents must investigate their surroundings to discover the best policies through trial and error. However, in situations where incentives are few or where mistakes are undesirable, and safety is a top priority, exploration becomes difficult [30], [31].

A different approach might evaluate abstract problems like considering systems that might outsmart us or the ethical considerations on the usage of AI, for example, using the Generative Adversarial Network(GAN) such as the "deepfake" to spread disinformation [32] using Deep Neural Networks (DNN), or the legal definition and usage of AI. Still, it is out of this study's scope.

This paper attempts to contribute to this field by offering a thorough analysis of the state of AI safety and trustworthiness attributes shown in Figure 1, talking about its difficulties, and offering potential research directions for the future.

The main contribution of this paper is that we answered the following research questions.

1) *How can safety be ensured through the design and implementation of learning techniques?*
2) *What are the main areas of investigation for AI safety?*
3) *What are the main concerns of safe AI, and how have these safety concerns in the previous literature been handled?*

The rest of this systematic review is organized as follows. In section II we describe some preliminary definitions. In section III, we compare our review to related reviews. In section IV, we present the research methodology to get the relevant studies. Section V presents and analyses the results of this review. In VI, we describe a discussion of this review, and finally, we conclude the safety of AI and future directions in section VII.

II. BACKGROUND

This background section thoroughly reviews AI safety and analyzes the safety issues raised by various learning techniques. It also offers fundamental definitions and the importance of further research on AI safety and trustworthiness.

A. AI TRUSTWORTHINESS AND SAFETY

In today's technological world, AI trustworthiness refers to the systems' resilience to adversarial attacks and their robustness, reliability, safety, explainability, and fairness.

AI safety has changed dramatically since the advent of AI and machine learning technologies. The need for a distinct field devoted to the safety of these technologies grew increasingly evident as they progressed. AI safety is now recognized as a crucial area for research due to the numerous complex issues it raises [33].

In artificial intelligence, robustness describes a system's capacity to withstand adverse circumstances or unexpected inputs and continue to function [34]. It denotes AI systems' capacity to tolerate challenging circumstances, such as threats to digital security. Robust AI systems can offer several benefits, including enhanced functionality, increased defense against adversarial attacks, and a lower chance of system failures. When presented with unexpected test samples from a distribution different from training, most modern ML algorithms classify them incorrectly despite their high degree of confidence; this epistemic uncertainty (unknown unknowns) can have disastrous safety repercussions [35]. Robustness is also referred to the ability of an AI system to handle challenging circumstances [36], [37], [38], [39], [40].

To ensure the robustness of complex AI systems, there are some of these challenges, such as mitigating bias and injustice in AI decision-making, making AI decisions transparent and explainable, and defending against adversarial attacks and out-of-distribution samples [41], [42], [43]. On the other hand, Explainable AI (XAI) is a collection of procedures and techniques that enable human users to understand and trust the output and outcomes produced by machine learning algorithms [43]. It's employed to explain an AI model, its anticipated effects, and any possible biases. Explainable AI facilitates a deeper understanding of artificial intelligence models and their decisions for developers and users.

In addition, reliability is essential in AI systems to lower the likelihood of system failures. Explainability helps people trust AI systems by making decision-making processes more understandable. Reliability is ensured by an understanding of the system's principles, and transparent AI systems make it easier to detect anomalies and errors, gradually increasing the system's robustness.

Model-agnostic methods provide distributional or indicator-based reliability estimates; for instance, they use sensitivity analysis or local cross-validation [39]. Model-agnostic methods do not depend on the internal structure or state of the trained model and can be used with any model. Although model-agnostic techniques are generally slower than model-specific techniques, they can apply to a wider range of models and architectures.

On the contrary, model-specific approaches depend on having access to the internal makeup and state of the trained model and are built to operate with specific models. These techniques are often more effective than model-agnostic techniques, although they are only applicable to specific models and architectural kinds [44]; this techniques are often known for their usage of probabilistic interpretations for reliability [45].

Moreover, for the system to be fault tolerant and to remove the negative effects of the malfunctioning parts on the system's regular operation, the locations of the faulty components and the severity of the malfunctions described by their types, shapes, and sizes must be known. Comprehending failure types and processes in the context of AI safety is essential. It aids in locating possible weak areas in AI systems and creating plans to stop them from failing, improving the security and dependability of these systems.

To guarantee AI safety, several norms and regulations have also been set. However, these methods and standards do have some drawbacks. For example, many safety analysis techniques currently in use are inapplicable to neural network-based systems. This emphasizes the need for more study in AI safety, which is still developing and requires much effort. Verification and validation, or V&V, are essential procedures that guarantee the system's trustworthiness [46]. Verification entails confirming that a system is reliable and secure, spotting any potential biases or flaws, and ensuring the system satisfies predetermined standards. Validation assesses a system at the beginning or conclusion of the development process to see if it meets the requirements as stated. In safety-critical sectors, where biased or inaccurate judgments can have dire repercussions, V&V is especially crucial. By offering proof that the AI-enabled system has undergone extensive testing and satisfies the specified standards, they contribute to developing confidence in the system. V&V is crucial to AI to guarantee accurate, robust, and trustworthy systems, minimize possible harm, and increase user trust.

Finally, numerous safety-critical systems are complemented with a Safety Instrumented System (SIS) to perform certain control functions and ensure the process continues to operate safely in the event of dangerous or hazardous situations; it is referred to as fail-safe operations [16], [47]. The fail-safe principle is one of the most crucial safety engineering techniques. It is critical to have a measure for the prediction's uncertainty. If this is comparatively high, the machine can ask for additional human verification [48].

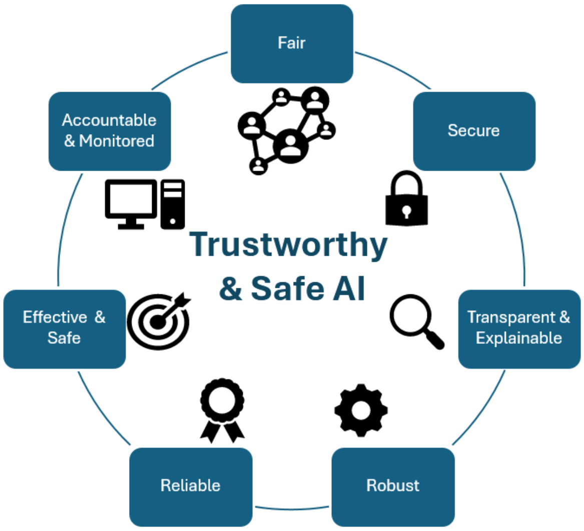

FIGURE 1. AI trustworthiness components.

B. AUTONOMY, CONTROL, AND INTERACTION

AI systems' autonomy or automation level defines their freedom from human oversight and control. It establishes the scope of control options available to the operator and the amount of system behavior information. Automation evaluation necessitates considering both the autonomy and human support of the application [48]. Control is a different aspect of a system that safety experts are interested in. Agent control can be defined as the dependability of an agent's actions and decisions to assess control concerning particular behavioral traits, such as accomplishing a task or abstaining from "risky" action [49].

An AI system's level of **autonomy** is decided by how it interacts with a human user to make decisions and who is ultimately accountable for those decisions. For example, full autonomy describes an AI system that makes decisions without human involvement. The levels of autonomy influence the structure of capacity, expectation, and safety in autonomous systems. For example, a railway scheduling system is an example of a fully autonomous AI system, while partial automation with a is also possible [50], [51]. In addition, the participation of humans in the AI system's decision-making process can range from an AI system making decisions with ongoing human oversight to an AI system offering suggestions to a human who then makes the final decision as shown in table 1.

The main variations between levels relate to the decision-making process between an AI system and a human user and who is ultimately accountable for the choice and control [50]. Since AI systems are growing more sophisticated and potent, it's critical to be able to watch out for any warning indications of potential issues and intervene to stop AI systems from doing damage.

TABLE 1. AI autonomy levels: The five levels.

Source	Description
0	No AI Autonomy
1	AI makes suggestions for humans (GPS guidance)
2	AI makes decisions under constant human supervision
3	AI makes decisions without constant human supervision
4	Full Autonomy (self-driving car with no driving cockpit)

C. LEARNING TECHNIQUE SAFETY

A crucial implementation of reliable and robust artificial intelligence systems is to ensure the safety of AI learning methods, including supervised, unsupervised, and reinforcement learning. This section explores the particular safety issues and factors to be taken into account for each of these learning paradigms, emphasizing their importance within the larger framework of AI safety.

In supervised learning, a model is trained with a labeled dataset in which the considered ‘‘correct’’ responses are known [52]. The training data quality and the model's capacity to generalize to new data are the main sources of safety concerns in supervised learning. The model may learn wrong patterns if the training data is of poor quality or contains errors, which could result in subpar performance or unexpected behavior [53]. Furthermore, an overly complex model may overfit the training set and perform poorly on newly discovered data [54]. Therefore, there are three crucial safety factors in supervised learning to assure the safety of the model: the data is of high quality, using an acceptable model complexity, and securely deploying the model.

On the other hand, in unsupervised learning, patterns are discovered in data without the assistance of labels or previous results [55]. Since there is no reliable means to compare findings to ground truth offline and conducting an online evaluation can be too dangerous, this presents a greater risk than supervised learning. The quality of the input data and the assumptions made by the algorithm have a direct impact on the safety of the model, such as clustering methods; for example, the misleading clusters may cause issues in situations where safety is a concern, such as in finance or healthcare [56].

In addition, reinforcement learning involves an agent learning to make decisions by interacting with an environment. Safety in reinforcement learning is a crucial topic, especially when these systems are implemented in the real world, such as in autonomous driving or robotics [57]. Safe reinforcement learning can be described as the process of learning rules that maximize the expectation of the return in challenges in which it is vital to assure reasonable system performance and respect safety limitations during the learning and deployment processes [58]. This involves refraining from activities that can negatively affect or result in penalties. The challenge is enabling the agent to explore the surroundings and pick up valuable skills while avoiding hazardous situations.

III. RELATED WORK

In this section, we delve into the existing literature reviews within the field of AI safety. Our aim is to offer a broad overview of the discussions and conclusions that have emerged in this area.

Through our analysis of these literature reviews, our intent is to underscore the distinctive contributions of our research in advancing secure and reliable AI. Subsequently, we will expound upon how our work either complements, challenges, or diverges from the conclusions drawn in these reviews.

Ashmore et al. [23] conduct a comprehensive overview of assurance techniques for ML. Their study highlights the significance of proving that ML is safe for its intended purpose and throughout its life cycle. Autonomy and human-AI interaction are not covered; instead, it concentrates on the iterative process of creating machine learning components for safety-critical systems.

Tambon et al. [59] presents organized, systematic literature on the certification of safety-critical systems based on ML. It addresses the difficulties and solutions in the literature for certifying machine learning systems in industries like automotive or aviation, where conventional certification methods are inapplicable. However, this review doesn't cover many attributes as we did in our review, like fairness, bias, and autonomy levels.

Myllyaho et al. highlights how crucial reliability is for AI systems, especially in light of the growing popularity of machine-learning methods that don't require extensive domain expertise. The study focused on the various validation techniques and combined them into a taxonomy that included expert opinion, simulation, model-centered validation, and trial. However, they didn't discuss the explainability and interpretability as we did in this review.

Vidot et al. [61] tackled the trustworthiness of ML and discussed its robustness and explainability. However, their survey doesn't cover the out-of-distribution samples and how to deal with them, which is a crucial part of ML certification.

Hamon et al. [62] addresses the difficulties in explaining AI models, especially in high-risk automated decision-making systems. It draws attention to the challenge of establishing unambiguous causal relationships between conclusions made using current machine learning techniques, particularly deep learning. However, a notable limitation of this article is that it didn't discuss the evaluation of explainability, and there is a lack of a thorough framework for assessing the explainability of AI systems that includes both user-centric and technical validation techniques.

In addition, Albahri et al. [63] presents a thorough analysis of reliability and explainability in the healthcare industry, classifying the research into seven domains, addressing issues, and offering suggestions for further study. However, it does not address issues such as adversarial and poisoning attacks.

Also, Adadi and Berrada [64] addresses the explainability of AI and the integration of AI systems into society, and the essay explores the significance of improving the transparency and trustworthiness of AI systems. It does not, however, address the robustness and reliability of AI systems.

Moreover, He et al. [16] addresses the problems of security, safety, health of the system, and moral quandaries in Human-Centered AI for reliable robots and autonomous systems. However, it does not adequately address the explainability of AI in detail as we did. Moreover, the authors of [65] highlight three important themes: interpretable methods, model behavior explanation, and reinforcement of safe learning. The results highlight that AI approaches with great application potential are vital for industry practitioners and regulators in safety-sensitive domains. Nevertheless, their study does not cover poisoning attacks and autonomy in AI systems.

Finally, the authors of [22] offer a thorough examination of the risks associated with ML concerning cyber-physical systems, especially regarding safety-critical usage like autonomous driving. The study highlights the necessity of an all-encompassing strategy for safety engineering and ML-based certification, outlining possible risks at every stage of the ML life cycle. The authors do not, however, discuss how humans and AI interact with these systems.

IV. RESEARCH METHOLODY

In this research, we followed Petersen's guidelines for systematic review [66]. The design of our study is presented in the following sections, along with information on the papers' databases, search terms and strategy, and inclusion and exclusion attributes.

A. RESEARCH QUESTIONS

We will discuss and survey the current state of the art for AI and present an overview of AI safety by comprising studies published from January 2018 to March 2024. During the systematic review, we used the following list of research questions that were coded as a guide during the analysis:

1) RQ1: How can safety be ensured through the design and implementation of learning techniques?
Justification: by answering this research question, we aim to characterize the intensity of scientific interest in the safety of intelligent systems and provide a solid foundation to classify existing research on the safety of AI.
2) RQ2: What are the main areas of investigation for AI safety? Justification: by answering this research question, we aim to present the recent strategies to mitigate the risks and ensure the safety of AI systems.
3) RQ3: What are the main concerns of safe AI, and how have these safety concerns been handled in the literature that has already been published
Justification: by answering this research question, our objective is to profile the challenges regarding the safety of AI.

B. ELIGIBILITY CRITERIA

We followed the above mentioned Petersen guideline and the 'PRISMA' [67] as a guide for searching and extracting data for our systematic review [68], [69]. This review's format was determined by the most recent PRISMA checklist (PRISMA Checklist 2020) [67]. Multiple sources were used to obtain the papers that fit the review's criteria following PRISMA guidelines. A paper must discuss AI or related fields, for example, safety or AI-based systems, learning techniques, and the main quality attributes of the AI model that must be regarded as credible. The information should also have been included in conference proceedings, workshops, peer-reviewed journals, or symposiums.

We set the interval of the electronic search from January 2018 to March 2024 to keep this review updated with the most recent and pertinent studies in the field; this interval guarantees that the information reviewed is as current as possible. In addition, this period was chosen to reflect the speed at which our area is developing and the likelihood that more recent research will offer more precise and up-to-date information. Finally, the interval selected aligns with the PRISMA criteria, which call for a precisely specified and justified time range for the literature search.

C. INFORMATION SOURCES

To guarantee the best outcomes, several academic abstracts and citation databases for peer-reviewed literature were used to cover the diverse papers. We chose the relevant sources that address our main topic "AI Safety". We found that Scopus, IEEE, and Springer are sufficient for a comprehensive review [70], [71]. These three databases give users access to millions of documents and have sophisticated, powerful search tools that make it easy to conduct in-depth literature searches. Although ACM and Elsevier are valuable resources, they were not used in this specific research because of the study's particular focus and scope. Future research could certainly contemplate incorporating these databases for a more thorough coverage.

D. SEARCH KEYWORDS

Carefully chosen keywords were used to find pertinent articles. Several keywords were created and then narrowed

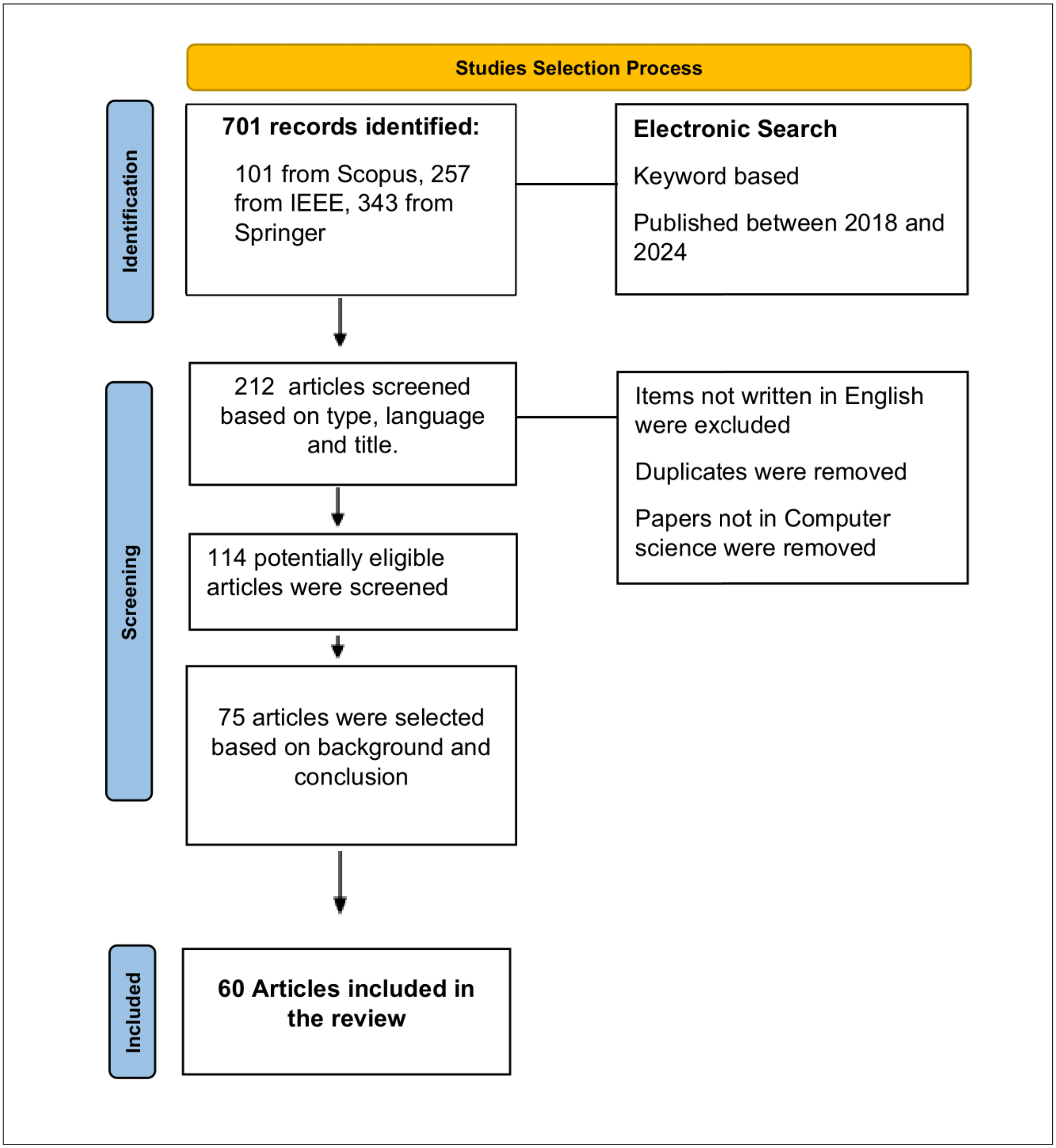
Studies Selection Process
Identification
701 records identified:
101 from Scopus, 257 from IEEE, 343 from Springer
Electronic Search
Keyword based
Published between 2018 and 2024
Screening
212 articles screened based on type, language and title.
Items not written in English were excluded
Duplicates were removed
Papers not in Computer science were removed
114 potentially eligible articles were screened
75 articles were selected based on background and conclusion
Included
60 Articles included in the review

FIGURE 2. Schematic showing the papers selection process.

down based on the research goals. Our search queries were created using ''Safety'' and ''Artificial Intelligence'' as main keywords. The queries shown in Table 2 are used with various syntaxes to adapt to the different rules set by each data source, even though they all have the same logical structure.

E. SELECTION PROCESS

The extracted data were narrowed down into different steps to choose which studies were pertinent for this analysis. We start by reviewing each document's title and abstract to determine whether or not it was appropriate to the study's subject. To extract only relevant articles that answer the research questions, particular inclusion (InC) and exclusion (ExC) criteria were set:

- `InC1`: The study needs to be an article, a conference paper, or a workshop.
- `InC2`: The study needs to be in the domain of Computer Science.

- InC3: The study needs to be a primary study.
- InC4: The study should discuss and address the Safety of AI.
- ExC1: The study is written in another language than English.
- ExC2: The research is duplicated.
- ExC3: The study is published as abstracts, editorials, and keynote speeches.

TABLE 2. Queries used to collect studies from different sources.

Source	Query
Springer	"artificial intelligence" and "safety" and "trustworthiness" and "safe" or "ai risk assessment" or "ai validation" or " ai assurance" From 2018 to 2024
Scopus	TITLE-ABS-KEY (”safe AI” OR ”safe artificial intelligence” OR ”AI safety” OR ”artificial intelligence safety”) AND TITLE-ABS-KEY (”machine learning” OR ”deep learning” OR ”reinforcement learning”)AND PUBYEAR >2017.
IEEE	("All Metadata":ARTIFICIAL INTELLIGENCE OR "All Metadata":INTELLIGENT OR "All Metadata":MACHINE LEARNING OR "All Metadata":DEEP LEARNING) AND ("All Metadata":SAFE) AND ("All Metadata":SAFETY) AND ("All Metadata":SAFE AI) AND ("All Metadata":SAFE ARTIFICIAL INTELLIGENCE)

F. SYSTEMATIC REVIEW EXECUTION

The earlier sections presented the steps to produce a systematic review. In this section the results of the systematic review after applying the inclusion and exclusion criteria where we collected 60 relevant studies presented in Table 3.

1) REVIEW PROTOCOL, DATA EXTRACTION, AND CODING

Following the PRISMA protocol, we outline our search and selection procedure as seen in Figure 2. Since we are looking for information on a specific topic, we used the title, abstract, and keywords in the default automatic search. The papers were examined for their titles, abstracts, conclusions, and full-text reading before being either included among the relevant articles or excluded as irrelevant for the review.

After applying the inclusions and exclusions using the appropriate filters on the aforementioned source sites and analyzing the title and abstract of each study, we ran the queries on different databases in March 2024, and 701 studies were returned (101 from Scopus, 257 from IEEE, and 343 from Springer) as a primary result.

2) RESULTS

The field of artificial intelligence research has seen tremendous growth in the last few years, with many studies delving into different aspects of this ever-evolving area. Figure 3 shows different article types included in this review. Most researchers nowadays are contributing to the development of the AI field, such as techniques and applications.

Moreover, figure 4 illustrates the increased number of selected publications on “Safe AI” from 2018 to 2023; it highlights the growing interest in and recognition of AI safety as a critical issue. In 2018, there was a shortage of research on AI safety, as seen by the small number of papers in this field. However, in the years that followed, there was a notable surge in study publications as the possible dangers of AI became more apparent, underscoring the increasing awareness of AI safety. Conversely, one possible explanation for the decreased number of publications during the last year could be the timing of data collection. The data for the previous year could not have included all the publications from that year they had been collected because of the publication process and duration. This can lead to a count that appears lower than in prior years. It’s crucial to remember that the data for the last year might not be comprehensive, but it could still be updated. The rise in scholarly contributions demonstrates efforts to establish proper safety standards amidst the rapid development of AI applications. The widespread discussion on AI safety, as indicated by the diversity of publication venues (Table 4), underscores the significance of robust safety standards for AI systems. This collective move by the research community emphasizes the importance of addressing safety concerns associated with AI.

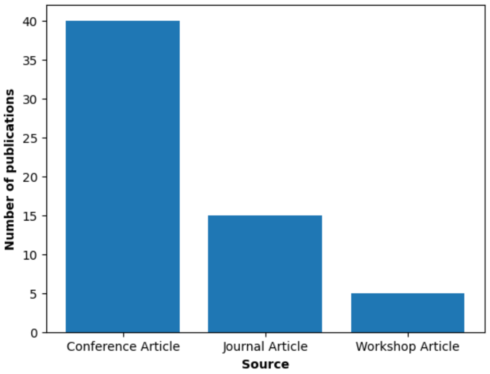

FIGURE 3. Total Number of included papers from each source.

Also, Figure 5 shows changes in the number of publications per year as a primary result for the electronic search using the queries presented in Table 2 (without using Inc and ExC). The results show an increasing number of publications per year between 2018 and 2023.

Still, the topic of AI safety has started attracting more researchers because of its importance to humanity, and there is a gap between the development of AI applications and the safety of these applications, which must be filled with proper standards to ensure the safety of AI systems. The complete

TABLE 3. A look-up table that classifies the acquired studies according to the safety concern.

Ref	Year	Safety Issue	Technique
[50]	2019	Autonomy Levels	Impact of different autonomy levels on safety assurance in AI-clinical systems .
[49]	2020	Safety Quantitative Metrics	They analyze AI Safety in terms of quantitative factors (Generality, Capability and Control).
[72]	2019	AI Failure Mechanisms	Outline different reasons for multi-agents failures in ML.
[73]	2021	Safety of RL agents	Forty-three safety properties to ensure the safety of deep RL agents in Atari games.
[74]	2022	XAI	SHapley Additive exPlanations.
[75]	2021	RL Safe Requirements	Transfer non-functional and functional requirements into shaped rewards to learn specification.
[76]	2021	Safe Autonomous systems	Human-in-the-loop methods to ensure safety of autonomous systems.
[77]	2021	Robustness	The effect of label noise on the robustness of deep detectors.
[78]	2019	RL Safety	Studying the safety of AI behavior by combining ML, game theory, and chaos theory.
[79]	2020	Fairness& Bias	Reduce algorithmic bias while hiding protected characteristics such race and gender.
[36]	2022	Robustness	CAISAR: an open-source platform for robustness and safety of AI systems.
[80]	2021	V & V methods	Detecting Adversarial examples to improve reliability of deep learning.
[39]	2022	Robustness&Reliability	Estimating the robustness of ML using statistical measures.
[81]	2019	Autonomy	A logical path towards safer operation of autonomous cyber-physical systems.
[38]	2022	Robustness	Adversarial patch attacks on multi-view detectors.
[82]	2021	RL Safety	Imposing formal safety restrictions on end-to-end policies in deep RL with visual inputs.
[83]	2020	OOD	OOD detection using generative models with multi modal prior distribution.
[84]	2019	Robustness Methods	Verify object recognition to assess the robustness of DNN against generic attacks.
[85]	2022	Poisoning Attack	Swarm optimization to calibrate DL algorithms in the presence of poisoning attacks.
[86]	2021	Robustness	Uncertainty handling using Bayesian Deep Learning control policies.
[87]	2022	XAI	Categorizing different types of XAI (Local vs global, model agnostic vs model specific, intrinsic vs post-hoc).
[37]	2022	Robustness	ARGAN: generative adversarial-based to defend against adversarial examples in ML.
[40]	2020	Adversarial attacks	Sparse regularization to strengthen DNN resistance against adversarial attacks.
[88]	2019	XAI	Counterfactual explanations to explain decision tree model in the finance applications.
[44]	2022	XAI	A lightweight method as a solution to the limitations of saliency mapping methods.
[89]	2022	Poisoning Attacks	Clustering-based label-flipping method against random flipping.
[90]	2019	Bias	Topological data analysis for bias visualization and mitigation.
[91]	2020	AI Safety Analysis	Proposing four AI system test strategies and introduces a multidimensional safety analysis checklist.
[35]	2020	Uncertainty	Fusion algorithm approach to manage epistemic uncertainty and ensure decision safety.
[13]	2020	Reliability	A novel safety argument framework for critical systems using DNN.
[92]	2018	Adversarial Attacks	Using RL to inject perturbations into adversarial samples to trick malware classifier.
[93]	2022	Safety Monitoring	RADICS system: a safety mechanism for machine learning-controlled Cyber-Physical Systems.
[6]	2021	XAI	Ensure XAI suitability between different AI techniques via explanatory, capabilities and requirements.
[15]	2023	Monitoring &Robustness	Involve leveraging coverage analysis to boost DNN confidence.
[16]	2022	AI Trustworthiness	Implementing trustworthy robots and autonomous systems regarding safety properties.
[17]	2020	AI Trustworthiness	Discussing the challenges and opportunities to design trustworthy robots and autonomous systems.
[94]	2021	Interpretability	A proof of concept for an interpretable deep neural network based on comprehensible requirements analysis.
[95]	2021	RL Safety	A deep RL approach to formal safety regulation implementation in end-to-end policies is presented.
[96]	2020	Adversarial Attacks	Examining the effect of poisoning attack on DNN performance.
[25]	2022	ML assurance	Assurance of the functionality and performance in ML-based systems.
[97]	2022	Bias Identification	Hierarchical clustering method to visualize bias in unlabeled datasets.
[14]	2022	ML Trustworthiness	Combining human and machine to improve ML trustworthiness.
[98]	2023	XAI And Trustworthiness	Ensuring AI safety via privacy protection, fairness, robustness, and accountability.
[99]	2022	Adversarial Training	Enhance robustness e by incorporating novel tool-assisted human attacks and conservative classifier thresholds.
[100]	2024	AI Trustworthiness	Defining KPIS constituting the notion of trustworthiness.
[101]	2020	Autonomy Levels	Determining the optimal level of decision-making autonomy for vehicles during the coordination process.
[102]	2023	Robustness	Using distribution-restrained softmax-loss function to enhance the robustness of deep learning models.
[103]	2023	Failure Modes	Fault Tree Analysis (FTA) and Failure Modes.
[59]	2022	ML Certification	Certification of ML-based safety-critical systems.
[104]	2021	Human-AI Interaction	Selecting training data based on ethical evaluations.
[105]	2023	Safety of AVs	Assessing the safety and security of AVs thorough simulation and hardware-in-the-loop testbed.
[106]	2023	AI Reliability	Qualification Process to ensure the reliability of AI/ML systems in military applications.
[107]	2023	OOD Detection	OOD detection in RL using bootstrapped ensembles and probabilistic dynamics models.
[108]	2023	Safety Certification	Safety certification of AI systems that uses regression models
[109]	2023	XAI	SAFE: a new method for counterfactual explanations for DNNs using saliency maps.
[110]	2023	Functional Safety	A comprehensive tutorial on the security and functional safety of AI in embedded systems.
[48]	2021	AI Trustworthiness	Highlighting the risks and principles involved, and emphasizes the need for safe and reliable AI.
[111]	2024	ML Safety Assurance	Safe MLOps Process for the continuous development and safety assurance of ML-based systems.
[47]	2023	Safe AI Development	The use of the Box-Jenkins framework for safe deployment and operation.
[112]	2021	Reliability	A framework for certifying the reliability of autonomous systems.

list of included studies is shown in Table 3, and the Venues are shown in Table 4.

V. LITERATURE REVIEW

This section addresses the three research questions about AI safety and its obstacles to describe the review's findings.

A. HOW CAN SAFETY BE ENSURED THROUGH THE DESIGN AND IMPLEMENTATION OF LEARNING TECHNIQUES?

To respond to the first research question, we discuss the safety of different learning techniques in this section, influenced by [113]. We take the ML life-cycle as a use case to present a taxonomy in Table 5.

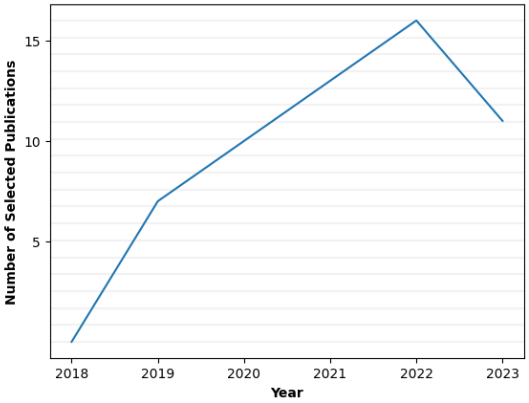

FIGURE 4. Selected papers regarding the topic "Safe AI".

TABLE 4. Number of obtained articles from each venue.

Source	#Venue
SAFECOMP2021	4
AI and Ethics	2
AISafety 2022	1
AISafety-SafeRL 2023	1
IEEE Access	2
SafeAI AAAI 2020	2
DASC 2020	1
ICSRS 2023	1
COMPSAC 2022	1
CVPRW 2022	1
VL/HCC 2022	1
ITC 2018	1
IEEE Transactions on Circuits and Systems II: Express Briefs	1
CVPRW 2021	1
IRCE 2020	1
IEEE Transactions on Cognitive and Developmental Systems	1
IEEE Transactions on Software Engineering	1
INDIN 2023	1
ICAA 2023	1
ICCPS 2022	1
RAMS 2024	1
ITSC 2023	1
Automated Software Engineering	1
Autonomous Agents and Multi-Agent Systems	1
Minds and Machines	1
International Conference on Human-Computer Interaction	1
ISR 2021	1
SAFECOMP 2020	1
AIMS 2018	1
NeurIPS	1
AAMAS	1
AI and Society	1
Big Data and Cognitive Computing	1
Proceedings AAMAS Conference	1
ICAA 2022	1
International Conference on Agents and Artificial Intelligence	1
Proceedings - Winter Simulation Conference	1
SAFEAI - AAAI 2020	1
51st Annual IEEE/IFIP DSN-S 2021	1
SAFECOMP 2022 Conference	1
QEST 2019	1
ATSIP 2022	1
HSCC 2021	1
Communications in Computer and Information Science	1
ACIT 2022	1
MIPRO 2022	1
ISVC	1
SAFEAI -AAAI 2019	1
Pattern Recognition	1
International Journal of Distributed Sensor Networks	1
IJCAI	1
HCII 2020	1
ICASSP	1
International Symposium on Leveraging Applications of Formal Methods	1

1) SAFETY OF LEARNING TECHNIQUE: RL CASE

Agents face a trade-off between exploration and exploitation in RL, where exploration can lead to better policies and rewards over time. RL enables systems to anticipate the effects of their actions and adapt to changing circumstances to ensure safety [28], [95], [121].

For that, Ritz et al. [75] suggests integrating both functional and non-functional requirements to help the agents learn to comply with the specification. For the reward function, the study demonstrates how the suggested method enables agents to manage intricate safety limitations and meet a predetermined set of requirements.

Furthermore, the authors of [78], [122] outline the main issues with AI safety for RL presented; these concerns are:

- Safe exploration: Can agents interact with their surroundings without harm?
- Resistance to shifts in distribution: Can agents indicate uncertainty about the applicability of their model to new kinds of data instead of operating under unsuitable models?
- Reducing undesirable side effects: Is it possible to design the agent's incentive system to avoid hurting its surroundings without covering every possible case in detail and explicitly?
- Steer clear of wire-heading and incentive hacking: Is it possible to stop agents from falsifying their observations to maximize their reward?
- Scalable management: Even with delayed input, can agents adapt their behavior and learn the right one?

The current safe RL techniques are either not formally guaranteed or based on erroneous assumptions about the reward structure and state space [82]. RL agents frequently behave erratically when put in circumstances outside their training environment [107], which can result in decreased performance or safety violations.

For that, Hernández-Orallo [49] suggests three criteria for analyzing AI safety: capability, generality, and control. They contend that these variables can influence one another and the risk that artificial intelligence systems pose. They employ a cutting-edge technique based on Agent Characteristic Curves (ACCs), which plot an agent's success probability across a range of task difficulty levels to assess competence and generality. Capability is represented by the area under the curve, while generality is represented by the slope. They characterize danger as the likelihood of slipping into a pit in a grid environment and control as the inverse of the entropy of the states the agent visits.

Hunt et al. [82] proposed a system called VSRL to achieve verifiably safe exploration and reward optimization, which

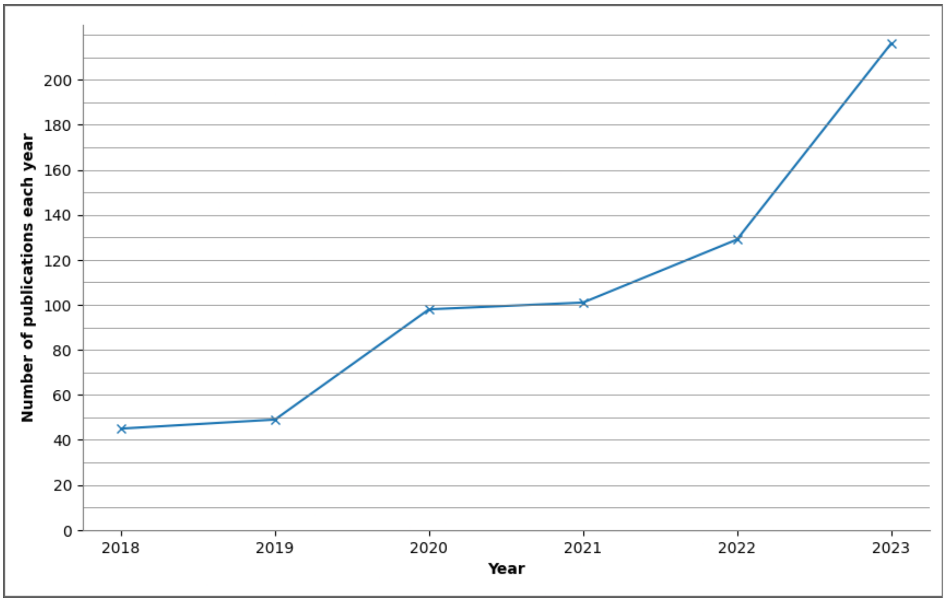

FIGURE 5. Published papers each year regarding the topic "Safe AI".

integrates object identification, automated reasoning, and deep RL.

Furthermore, Giacobbe et al. [73] provides a unique technique for evaluating and guaranteeing the security of deep reinforcement learning (DRL) agents in the complicated and unpredictable settings of Atari games. They establish forty-three safety properties for thirty-one games and assess how well five cutting-edge DRL algorithms work with them. They present a countermeasure that increases the safety of all agents over many attributes by using bounded prescience to protect the agents from risky acts.

2) SAFETY OF LEARNING TECHNIQUE: SUPERVISED-LEARNING CASE

Supervised learning is a subcategory of ML characterized by using labeled datasets to train algorithms that correctly identify data or predict outcomes [91]. For that, we present a taxonomy of ML life-cycle safety in Table 5, influenced by [113]. This taxonomy shows the safety of AI models by taking the ML life cycle as a use case. We discuss safety on three levels: data protection, safety in model development, and runtime safety. Indeed, those levels represent the life cycle of an ML model. The objective is to guarantee the safety requirements of AI models at every level and analyze the current approaches to help meet these requirements [23]. Ensuring the safety of each cycle is important to achieve the overall safety of any AI-based system. These attributes are linked directly to the life-cycle of any AI-based model, and the safety factors can be shown in Table 6.

a: SAFE DATA PROCESSING

The test and training data quality used to validate and train the model is guaranteed at this point [111]. It is a crucial phase for training and testing ML models, as it involves collecting, annotating, pre-processing, and augmenting data [38]. The quality and quantity of data directly impact model accuracy and robustness [36], [37], [38]. However, data processing can introduce hazards such as inadequate distribution, insufficient dataset size, and bias [90]. Inadequate distribution can lead to poor generalization and unexpected behavior [15], [83], [86]. At the same time, insufficient dataset size can overfit the model and fail to capture the complexity of the problem [16]. For that, several techniques are used to enrich the dataset, for example, using **geometric transformation** (rotation, crop, contrast enhancement, etc.) or **GAN** to generate or transform data [14], [37], [38], [97]. However, for those AI apps to become more trustworthy, the backend that handles the data collection and processing should consider the participant's data privacy as well as data integrity and authenticity [85]. **Data encryption** techniques can be used to ensure data privacy; through this process, the information is transformed from its original format into a different format or cipher text. For example, Watermarking allows an optical encryption system to conceal the presence of confidential data [116].

b: SAFE MODEL DEVELOPMENT

is selecting, designing, training, and tuning machine learning models based on available data. It aims to optimize a predefined objective function, which measures the error between model predictions and true labels [80]. However, model development can introduce hazards that compromise the safety of ML-based systems. These include incorrect objective function definition, inadequate performance measure, model complexity, and model uncertainty [76], [81]. The objective function defines the model's goal and evaluation, but if it's not aligned with the system's desired behavior

TABLE 5. Taxonomy of ML safety techniques.

Phase	Solution	Technique
Safe Data Processing	Data augmentation	Active learning[76, 114, 115]
		GAN [37, 97]
		Geometric Transformation [14, 38]
	Data Privacy	Data Encryption [116]
	Bias Detection	Regularization approach [37, 40, 77, 79, 86]
		Topological analysis [87, 90]
		Hierarchical clustering [97, 117]
Safe Model Design	Model Transparency (XAI, interpretability)	Visual representations [16, 40, 44, 87, 94]
		Counterfactual justifications [6, 88]
		Shapley Additive Explanations [74, 87]
		Saliency Maps [6, 44, 87, 94, 109]
	Robustness	Adversarial training [6, 15, 37, 39, 40, 59, 76, 79, 80, 84, 89, 92, 96]
		Transfer Learning [14, 77, 79, 90]
		Out-of-distribution[14, 15, 36, 76, 83, 86, 107, 118]
		Distance based prediction [13, 39, 89, 94]
		Model Extraction attack [119]
		Poisoning Attack [89, 96, 120]
	Fairness	Algorithmic Fairness [88]
Runtime Safety	Reliability	Fault diagnosis [16, 17]
		Local cross validation [39]
		Sensitivity analysis [39]
	Control	Autonomy [13, 16, 17, 25, 50, 81, 91]
	Monitoring	Anomaly detection [16, 17]

or safety constraints, it may lead to unsafe results [39]. Choosing an appropriate performance measure that reflects safety requirements and trade-offs is crucial for ensuring safety. Model uncertainty, arising from sources like noise in data or parameter uncertainty, can also affect the safety of ML-based systems by causing incorrect predictions or decisions [22], [77]. Moreover, model attacks, which involve malicious attempts to manipulate or degrade an ML model, can also affect the safety of ML-based systems [37], [40].

c: RUNTIME SAFETY

to make sure that a ML model operates as best it can in an actual setting, deployment is an essential procedure that requires multiple steps.

It entails monitoring the model's output over time and looking for notable deviations or changes. When a model is experiencing problems like concept drift a situation in which the data it is forecasting no longer matches the data it was trained on monitoring can help detect when a model may need to be retrained.

The most fundamental function of fault diagnosis is fault detection; it checks for system malfunctions or faults and establishes the time at which they occur [13], [84]. It can be applied to a decision-making issue in machine learning [16], [17], and more details about failure modes are discussed in b2.

3) SAFETY OF LEARNING TECHNIQUE: UNSUPERVISED LEARNING CASE

Unsupervised learning methods, which identify patterns in data without predefined labels, require careful implementation and verification for various reasons. The input data's accuracy directly impacts the results, making it crucial to work with clean, well-prepared data. Understanding the results can be difficult as the identified patterns may not correspond to known categories. Additionally, these methods may yield different outcomes with slight changes in data or parameters, underscoring the importance of conducting stability checks. Validation methods confirm the consistency of identified patterns with facts, evaluate the reliability of findings under varying circumstances, and prevent overfitting by identifying overly intricate patterns specific to the training data [112]. Therefore, while providing valuable insights, unsupervised learning demands careful application and thorough validation to ensure dependability and usefulness. More details about the validation methods are discussed in b1.

On the other hand, the validation of techniques requires careful consideration, mainly when dealing with **out-of-distribution (OOD)** inputs. OOD is another category of risky inputs [14], [15], [36], [76], [83], [86], [107], [118] including those that fall outside the training data distribution. Detecting OOD is crucial in safely deploying machine learning models, especially in unsupervised learning. It involves the model's ability to identify and handle input data that differs significantly from its training data. This is particularly important in unsupervised learning since the lack of labels can make it difficult for the model to interpret such data accurately. Hence, OOD detection mechanisms are commonly incorporated into unsupervised learning systems to improve their safety and dependability.

Haider et al. [107] suggest a novel method for OOD detection in reinforcement learning by characterizing OOD as severe perturbations of the Markov decision process (MDP) and presenting a predictive algorithm utilizing bootstrapped ensembles and probabilistic dynamics models.

In addition, **distance-based** prediction can be used to classify OOD. For example, each input and its neighbor should be the same if the distance between them is less than a specific number. This can put a limit on the generalization error and give people more trust in the DNN's reliability promise [13].

Rossolini et al. [15] proposed an approach for this issue: to use coverage analysis to boost deep neural network confidence. According to experimental results, the proposed approach successfully detects strong adversarial examples and OOD inputs.

B. WHAT ARE THE MAIN AREAS OF INVESTIGATION FOR AI SAFETY?

As artificial intelligence advances, protecting it becomes increasingly important. This section examines three primary areas of AI safety research: verification and validation methods, failure modes, autonomy, and interaction.

1) VERIFICATION AND VALIDATION METHODS

Verification and Validation (V&V) are procedures to ensure a product, service, or system meets requirements and specifications using different methods and various formality levels [112]. Verification evaluates if a system meets design specifications, while validation ensures it meets customer and stakeholder needs. In systems, validation involves acceptance and suitability with external customers, ensuring the system meets user operational needs. In the context of AI dynamic intelligent systems, these systems require non-conventional methods for verification and validation due to their constantly changing context. Traditional V&V methodologies focus on whether systems behave correctly, but AI V&V presents unique challenges. Significant obstacles are the inability to clearly describe correctness standards for system outputs and the difficulty in standardizing non-conventional approaches due to adaptive systems' changing context.

Box-Jenkins technique is a well-known framework employed to recognize and resolve engineering hazards in adjusting and validating AI models to ensure AI safety [47]. The Box-Jenkins technique for AI safety essentially aims to minimize methodological gaps in the engineering of these systems and ensure that AI systems fail safely, especially in critical systems [123]. This is especially crucial because of the possible harm that specific AI algorithms may cause to human integrity and their high predictive capacity. The Box-Jenkins framework has three primary phases for AI safety, as shown in Figure 6. The initial stage, identification, is preparing the data and choosing the suitable model to fit the relevant data. To ensure the model is built correctly and accurately represents the real-world system it is meant to imitate, the second phase, Estimation and Validation, estimates the parameters of the selected model and verifies the model's fit to the data. The last stage, Application, evaluates the chosen model's capacity for forecasting and applies it to actual situations. This iterative approach ensures AI systems' safe and efficient implementation, especially in crucial safety-critical applications. Moreover, Zhao et al. [80] highlights the shortcomings of existing testing methods that ignore the operational profile and may not increase software reliability. The authors introduce the idea of ''operational adversarial examples'' (AEs), which are more likely to arise in actual operations. They also recommend a novel testing methodology incorporating the operational profile to identify these AEs effectively.

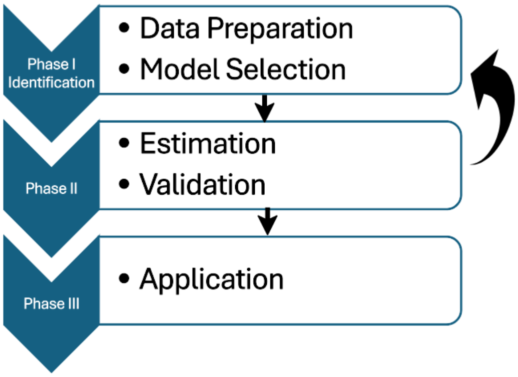

FIGURE 6. The Box-Jenkins framework for safe AI.

Finally, the authors of [13] present a safety framework for deep neural network-based critical systems, highlighting the significance of verification and validation (V&V) for safe operation. It uses V&V methodologies and operational data to forecast system reliability using Conservative Bayesian Inference (CBI). The decrease of generalization error, a crucial component of system dependability, is linked by the framework to several stages of DNN life-cycle activities.

2) FAILURE MODES

A fault is an abnormal state or defect causing failure in a component, device, or subsystem [124]. Prediction foresees future failures, while diagnosis identifies and assesses the component's fault [16]. **Fault diagnosis** is frequently used to track down, detect, and pinpoint defects to increase a system's reliability [125]. Fault detection, isolation, and identification are the three steps that make up fault diagnosis [126], [127]. The most fundamental function of fault diagnosis is fault detection, which checks for system malfunction or faults and establishes when they occur.

Manheim [72] discussed several failure modes for multi-agent systems where the failure modes become more complex. The study examines several multi-agent system failure scenarios, such as unintentional steering, misaligned adversaries, faulty coordination, input spoofing and filtering, and goal co-option. Compared to single-agent failures, these modes are less known and more complicated. Moreover Martinez et al. [103] discusses the use of reliability analysis methodologies like Fault Tree Analysis (FTA) and Failure Modes Effects and Criticality Analysis (FMECA) in analyzing AI systems' learning and adaptability, highlighting potential issues with system reactions to malfunctions. It provides insights and encourages further research in AI safety, potentially influencing industry standards.

3) EXPLORING AUTONOMY, OVERSIGHT, AND SAFETY IN AI SYSTEMS

Human-AI interaction is studying and creating artificial intelligence systems that can communicate and cooperate. It is becoming popular, emphasizing the importance of designing systems that align with human values and work well with humans [14], [16]. Hence, the safety of such AI systems must be ensured to minimize any risk that threatens the human. Human-centered AI refers to the development of AI systems that prioritize human control and autonomy while also leveraging the capabilities of AI to enhance human performance [128]. With the use of machine-based methods, humans can directly complete activities that are challenging for computers in the pipeline and give training data for ML applications [129].

The human-centered AI framework attempts to accomplish these objectives by accounting for high degrees of automation and human control. This strategy can aid in preventing mishaps and errors brought on by excessive automation or human control [130]. There is rising interest in researching more sophisticated learning strategies like **active learning**, which enables AI models to learn more successfully by actively choosing the data they learn from [72] and [76]. Active learning is a semi-supervised technique in which only a part or subset of the training data is labeled [76], [114]. When the model determines that a particular data point is of interest, an interactive query is made to a human to label the data points from the unlabelled set.

Hagendorff [104] emphasizes the influence of social and psychological backgrounds on AI behavior to develop ethically sound and socially conscious AI applications and suggests a selective use of training data based on ethical judgments. These systems must be made safe by defining the conditions under which they are intended to operate, identifying potential dangers, and guarding against AI model and system hardware failures. Rajendran et al. define in [76] the ways that people can contribute to the learning process: Learning from demonstration, intervention, and evaluation are methods used by computers to learn from human actions.

In active learning, the premise is that a person will always be able to identify the correct label for a challenging data point. Some data points, though, might be difficult for human labelers to interpret. For example, it is assumed that the human did not make any mistakes while executing the demonstrations and that the demonstration data is a good sign of a safe policy. Although there is a chance that there may be ambiguity between various human evaluators, it is assumed that the human provides the proper reward for the right action [76].

Furthermore, Bravo-Rocca et al. [14] proposed a multi-agent system in which a human agent works with the other three agents to categorize anomalous events using the human-in-the-loop technique. This human involvement enables more precise data labeling and can raise the ML model's credibility. Human-in-the-loop systems may be safer than fully autonomous ones [47]. This approach can help make the model more trustworthy, safe, and reliable in real-world scenarios where every choice has some responsibility. This can lower the possibility of inaccurate forecasts and increase safety in critical applications [14].

Moreover, Zhang and Cho [131] proposed the "Safe DAgger" method, which computes the difference between the expert-level actions of humans and those of agents and uses this as a decision measure for querying. If there is a slight difference between the actions, the agent action is sampled; if there is a significant difference, the oracle is consulted to determine the proper action. The difficulty in defining the metric and threshold choices for each application and circumstance is a downside of this method [131].

Finally, Mariani and Zambonelli [101] addresses autonomous vehicle coordination solutions, emphasizing the difficulties in choosing the right action strategies given the traffic situation in different scenarios. The result is that resolving these issues and considering ethical dilemmas in vehicle coordination is necessary to ensure the safe and conflict-free movement of autonomous cars.

TABLE 6. AI safety issues and examples.

Safety Issue	Examples of this AI issue
Adversarial-Attacks	Adversarial samples, data poisoning.
Reliability	System acts as intended for a specific period.
Robustness	How will the system react to new or unseen samples?
Explainability	How to explain ML model's behavior in human terms.
Bias	AI's decisions are based on race or gender.
Fairness	Unbiased decisions for all users.
Control	Take control of AI-based system in case of urgency.
Interpretability	Interpreting the model's features and weights.
Transparency	Understanding the model's inner parts.
Security & Privacy	Secure data and code from malicious attacks.

C. WHAT ARE THE POTENTIAL CHALLENGES OF SAFE AI?

In this section, we try to discuss the main challenges regarding the safety of AI models, mainly the quality attributes such as robustness, reliability, and explainability. The main safety factors for any AI model can be shown in Table 7.

1) ADVERSARIAL ATTACKS

Since ML models are the primary source of cognitive cyber security, they might be attacked. ML algorithms are a program or a function within a program that could be compromised [16]. Hackers could add training examples or replace them with adversarial samples in which hackers insert perturbation samples that lead an ML model to make an inaccurate prediction. In addition to that, hackers can change the code of the model.

Therefore, as a result, ML model protection should address two issues: securing the ML model's code and finding a solution to the adversary fitting. When designing the system, it is necessary to look into the security and robustness of ML models [17], [118].

TABLE 7. AI safety properties.

SafetyFactor	Reference
Robustness	[102], [36], [80], [39], [38], [86], [84], [37], [102]
Reliability	[92], [40], [89], [35], [91], [13], [112], [106]
Bias & Fairness	[79], [90]
Transparency	[44], [87], [88], [13], [6], [74], [98], [109]
OOD	[76], [83], [86], [15], [107], [36], [14], [118]
Trustworthiness	[90], [97], [14], [16], [79], [98], [48], [104]
Adversarial Attacks	[16], [118]
Poisoning Attacks	[85], [89], [96]

The term "adversarial samples" refers to situations where small feature perturbations can lead an ML model to predict incorrectly. Adversarial examples expose ML models to threats [15], [92]. For this purpose, **adversarial training** ensures AI safety against adversarial attacks. These attacks entail adding meticulously crafted perturbations to input data nearly indiscernible to humans but can lead models to produce inaccurate predictions. In general, because it forces the model to behave unexpectedly, enabling the observation of how it responds to worst-case input where a single mistake might have disastrous consequences [16], [99].

Adversarial robustness is the ability of an ML model to withstand adversarial attacks. In the context of adversarial attacks, model robustness is defined as the "average magnitude of the minimal adversarial perturbation over many samples" [39].

Most studies in this area concentrate on creating ML models that are resistant to several adversarial attacks. The following are some examples of the traditional adversarial defense strategies mentioned in [15], [92], and [117] such as adversarial training, randomization techniques to lessen the impact of adversarial samples, and strategies for proving defenses based on well-defined attacks.

Tarchoun et al. [38] have studied the resistance of multi-view detection to recent adversarial patch threats by using perspective geometric transforms; the authors suggest an evaluation framework in which an adversarial patch is trained against a single view of a multi-view data set and then applied to the other views of the data-set.

A recent study has shown that DNNs aren't resilient to several attacks. The DNN cannot classify the adversarial instances, which contain images that are strikingly similar to the ones that were correctly categorized.

In addition, Mziou Sallami et al. [84] focuses on formal methods for NN-based object recognition systems and presents a novel method for evaluating the robustness of an NN-based image classifier. The authors offer a thorough method for assessing object recognition systems using Abstract Interpretation theory, which may be used to assess a DNN perception system's resilience against more general attacks.

Choi et al. [37] presents "Adversarially Robust Generative Adversarial Networks" (ARGAN). This novel GAN-based protection technique transforms the input data into ML models using a two-step transformation architecture. The target deep neural network model's vulnerability to adversarial instances is reflected in the generator model, which also optimizes its parameter values for a joint loss function. The authors show how ARGAN preserves the acceptable accuracy for legitimate input data while preserving the target deep neural network model's robustness against hostile samples.

Finally, Wang et al. suggest 'DRSL' [102], a technique that improves deep learning models' resistance to adversarial attacks without causing appreciable time consumption or accuracy loss. It suppresses distribution diversity in softmax values for non-real label samples.

2) POISONING ATTACK

ML models may become corrupted due to the serious issue of data set poisoning or **poisoning attack**. Data set poisoning involves the attacker inserting false or inaccurately labeled data into the data set, thus turning the accurate labels into incorrect ones. Label flipping attack can be done by replacing some true samples in a dataset to alternate labels selected from the dataset in place of the true labels of the features vectors to make the model predict the income samples according to the chosen labels and targets [96].

Maabreh [89] propose a clustering-based label-flipping attack method and evaluate its effectiveness on several commonly used ML algorithms. The primary goal is to generate poisoned training samples that affect classifier accuracy as they go through outlier detectors. Their approach shows good results when using ML techniques such as K-Nearest-Neighbour (KNN) or Binary Decision Tree (BDT), but it fails with Random Forest (RF) and DNN.

On the other hand, Maabreh et al. [85] examined how Particle Swarm Optimization (PSO) can enhance these models' performance in the presence of poisoned data and evaluated the performance of deep learning models under various poisoning rates. They expressed concern that PSO might mask or hide the effects of poison, which might result in inaccurate learning during the later stages of improving the model using more recent data.

3) BIAS AND FARINESS

AI-based systems aim for fairness in predictions, ensuring they are unbiased and not based on factors like race or gender. However, unjust AI-based systems can lead to ethical issues and financial losses [13], [36], [79], [87], [90], [97]. Three main approaches have been developed to address AI bias: pre-processing, in-processing, and post-processing. Pre-processing targets data bias, while in-processing removes algorithmic bias by modifying the learning algorithm with constraints. Post-processing techniques minimize bias in predictions after training a machine learning model. The difference between these categories is shown in Table 8.

Kamishima et al. [132] proposed a **regularization approach** to resolve algorithmic bias based on analyzed unfairness. Moreover, Jaipuria et al. [97] proposes a unique **hierarchical clustering** approach that makes use of deep perceptual properties. Using deep perceptual characteristics and similarity metrics, deepPIC is a technique that organizes unlabeled picture files into semantically meaningful categories. It can assist in visualizing and comprehending the kind and degree of bias present in various datasets. One drawback is that deepPIC might not be able to discriminate between nearby semantically comparable clusters, particularly in cases when the clusters are small. Another constraint is that deepPIC uses the computationally costly scaled Learned Perceptual Image Patch Similarity (LPIPS) score to measure perceptual similarity between image pairs. A further drawback is that deepPIC extracts feature using generic backbones trained on ImageNet, which might not capture the task-specific bias for various vision tasks [97].

Furthermore, Srinivasan and Chander [90] suggests a method based on **topological data analysis** to detect biases before the use of a bias mitigation algorithm. Their method employs persistence homology, a topological data analysis technique, to identify and measure bias resulting from various attributes in a dataset; it can identify clusters, holes, and voids at different spatial resolutions.

In addition, Kim and Cho proposes an unbiased information bottleneck method with adversarial learning to achieve fair representation. It achieves the highest performance in reducing the algorithmic bias on various datasets.

Finally, AI systems are educated on human-generated data, and this data may contain biases. It is crucial to be able to keep an eye out for bias in AI systems and act to eliminate any biases that are discovered. For that, Zhao et al. [133] introduced a customized probabilistic measure for **bias monitoring**, per their definitions, by using life-cycle activities to gain prior knowledge about the measurements. After that, statistical inference is performed under continuous observation to illustrate and comprehend bias in unlabeled and unstructured datasets.

TABLE 8. Processing time for different biases types.

	Pre-processing	In-processing	Post-hoc
Data Bias	✓	-	✓
Algorithmic Bias	-	✓	✓

4) RELIABILITY

The term reliability refers to the continuation of accurate service. It includes both the supply of services and the accuracy of the ML model's output, or the continuity of accurate service [13], [16], [38], [39], [44], [83], [94].

The context of ML-model-driven systems includes both the provision of services and the accuracy of the ML model response and correctness. Model-agnostic or model-specific techniques are the two categories into which existing reliability estimate methods may be divided.

Akram et al. [39] proposed a metric known as StaDRe (Statistical Distance for Reliability). They estimate the reliability of machine learning forecasting techniques for time series data. It evaluates the effectiveness of the machine learning mode and detects distributional shifts using statistical distance measures based on the Empirical Cumulative Distribution Function (ECDF).

5) ROBUSTNESS

Robustness in machine learning refers to a model's capacity to maintain its performance despite varying inputs from its training data [59]. Next, the term "model robustness" describes how much a model's performance varies between training and new data [100].

The ML literature evaluates a variety of model resilience properties, including robustness against adversarial assaults and robustness against data-set change, as well as robustness against OOD as discussed in a3. Arnez et al. [86] suggests a technique to take uncertainty into account while generating Bayesian Deep Learning control strategies; their preliminary research demonstrates that the suggested approach strengthens the navigation policy's robustness in OOD circumstances.

In addition, Kshetry and Varshney [35] proposes a technique for reducing epistemic uncertainty and maintaining the safety of decisions by fusing algorithmic knowledge with physical system information using data-driven trained models. Finally, Kamoi and Kobayashi [83] proposed another method that can be used by applying a multi-modal prior distributions model, which may reduce OOD likelihoods and serve as an OOD detector.

On the other hand, and to increase the robustness of AI systems, **transfer learning** is the act of transferring knowledge from one source task to help them accomplish another target activity more effectively [134]. Transfer learning offers a quick and easy method for creating effective ML models, especially when a lack of training data or computing power is available to complete the target job. This leads to robustness enhancement against new unseen cases.

The goal of adversarial robust transfer learning is to increase the robustness and safety of a target model by transferring its robustness from the source model. A simple transfer learning technique is used to retrain the new model using the weights of the existing model. This results in a more robust model once the new enhanced data is added to the training set. [14].

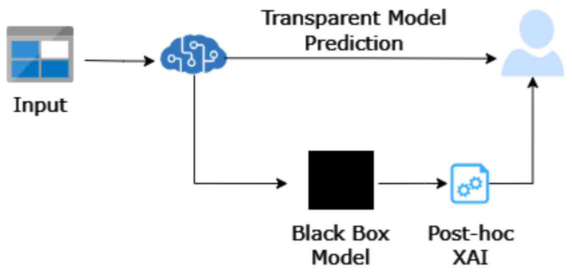

FIGURE 7. Types of XAI: Transparent vs blackbox.

6) TRANSPARENCY

XAI is a set of processes and methods that enable human users to understand and believe the output and results produced by ML algorithms [6], [44], [64], [86], [87], [88], [98], [135]. It discusses an AI model's expected effects and potential biases and allows people to comprehend predictions or judgments.

Interpretability can be global or local, with global approaches explaining how a model produces predictions holistically, while local methods focus on specific data samples. There are two approaches to interpreting an AI system: model agnostic and model specific, as shown in Figure 7. Using agnostic methods separates the ML black boxes' from explanations; this post-hoc interpretation occurs after the model is trained and is unrelated to its core design. Other models are transparent by nature and simple to understand intrinsically. Thanks to their underlying representations, such as the K-Nearest Neighbors, these models benefit from immediately supporting explanations. Many studies discussed the XAI; for instance, **counterfactual justifications** is used to describe how a prediction's outcome would have been different if particular characteristics had different values [6], [88], [109].

In addition, Sokol and Flach [88] used class-contradictory counterfactual assertions to explain decision tree models trained on two different data sets. These justifications can be used to find security and privacy flaws, test the underlying model, fix them, and eventually help make an AI system safer by identifying undesirable or dangerous behavior.

Cooper et al. [44] propose a new method based on **saliency maps** for clarifying and comprehending the predictions provided by AI systems. Saliency mapping can assist in making the inner workings of an AI system more apparent and accessible to humans by producing **visual representations** of the parts of input most crucial in influencing a model's output. According to the authors, their method may be helpful in various AI applications where explainability is crucial, but present methods are either too slow or have limited applicability. Also, Samadi et al. [109] propose 'SAFE', which guides the creation of more precise and relevant counterfactual explanations for DNNs in autonomous driving systems using saliency maps.

Local Interpretable Model-agnostic Explanations (LIME) explain models' predictions. It is employed to elucidate specific predictions made by machine learning models [59], [98], [100]. Users may learn how the model acts and which features have a major impact on the output by examining changes in predictions after perturbing input data points. LIME facilitates the creation of more understandable and transparent AI models, which aids in creating reliable AI systems.

On the other hand, Buczak et al. [74] uses **Shapley Additive Explanations (SHAP)** to explain specific model predictions. SHAP is based on ideas from cooperative game theory; each feature is given a value indicating how important it is for a specific prediction. A comparison of these different techniques is presented in Table 9.

7) MONITORING

As AI systems become more advanced, monitoring potential issues and intervening to prevent damage is crucial. Observability, or monitoring, is essential for AI-based systems to align with human values and ensure safe operation [91].

For that, Dementyeva et al. [93] uses the 'RADICS' system, which consists of black and white box monitors, to guarantee the security of cyber-physical systems managed by machine learning algorithms. If problems are identified, the system enters a safe mode, offering a thorough safety check for the decisions and results.

On the other hand, AI systems are educated on human-generated data, and this data may contain biases. It is crucial

TABLE 9. Classification of different XAI methods.

Method	Analysis Type	Type	Technique
Saliency Maps [6]	Local	Model Specific	Draw attention to the aspects of an AI model's input that have a big impact on the predictions made by the model.
SHAP [74]	Local	Model Agnostic	It determines each feature's marginal impact on the model's result, which allows it to assign each feature's contribution to the prediction.
Counterfactual Methods [88]	Global	Model Agnostic	Explain how changing a model's inputs might affect its outcome.
HiPe [44]	Local	Model Agnostic	Iteratively concentrates on the most significant regions with increasing resolution to provide robust saliency maps and interpretable explanations for model predictions.
Visualizing Methods	Global	Model Agnostic	Visualizing approaches provide heatmaps or other visual representations that show how input features affect the model's prediction.
LIME [59, 98, 100]	Local	Model Agnostic	It approximates the behavior of the complicated model and produces a simpler, more interpretable model, which produces explanations that are particular to individual cases.

to be able to keep an eye out for bias in AI systems and act to eliminate any biases that are discovered.

Zhao et al. [133] introduced a customized probabilistic measure for **bias monitoring**, per their definitions, by using life-cycle activities to gain prior knowledge about the measurements. After that, statistical inference was performed under continuous observation to illustrate and comprehend bias in unlabeled and unstructured datasets.

VI. DISCUSSION AND CHALLENGES

Researchers and engineers have employed a rigorous process to build diverse AI systems, safely handle failures, and continuously monitor their health and performance despite challenges.

One of the biggest challenges facing AI right now is its reliability. As time and technology advance, it becomes clearer that these systems must be reliable to attain the production plateau. A reliable model consistently predicts correctly and offers high confidence in its findings.

By incorporating human intelligence and judgment, humans would prevent the system from investigating unrelated conditions and states [76]. The difficulty is in guaranteeing AI's reliability in safety-critical systems, where malfunctions could have serious and disastrous repercussions. Another problem regarding this technique is the size of the training data which could be large which means more human effort to complete the labeling or demonstration.

Furthermore, there are various "schools" and motivations. Some experts might view the issue of AI safety as a variation of a classic engineering challenge in critical systems that includes some "cognitive" elements but primarily focuses on some attributes (robustness, reliability, etc.).

However, a different perspective might examine abstract problems like interpretability or the black-box problem. As AI systems become more complex, it becomes more challenging to comprehend how they operate and forecast their behavior in various scenarios. Hence, the complexity of AI systems will prevent us as humans from understanding what is happening inside the models because AI systems are frequently opaque, and it can be challenging to comprehend how they make decisions. For that, XAI tries to explain the decisions made by AI, but this comes with a cost in terms of performance, as shown in Figure 8.

As known to the "black box" theory of ML, the algorithm's decision-making process is opaque to even its creators [64]. It is crucial to accurately explain AI systems' predictions in a manner that is clear to people when these systems are utilized for more intricate and risky activities. This is essential for fostering confidence in AI systems and guaranteeing secure deployment. XAI is a set of procedures for comprehending and justifying the choices made by AI systems [44].

As a result, both XAI and interpretability refer to the capacity of an AI system user to comprehend its decisions; depending on the specific area of the problem (such as the medical domain, the economic domain, etc.), the manner and style of interpretation may change. The machine model of knowledge can approach the human model through a variety of processes. By enabling people to comprehend how AI is making decisions, the field of "XAI" tries to overcome these problems. XAI contributes to AI Safety when used in autonomous systems because it shows how decisions are made so humans can understand the system comprehensively.

However, one major gap in the XAI techniques is the evaluation of XAI approaches. Although XAI aims to increase the transparency of AI systems, evaluating the success of these justifications is still difficult. The main reason is that interpretability is subjective; what makes sense to one person may not make sense to another. Furthermore, comparing and contrasting various techniques is challenging because there are no established standards and benchmarks for assessing XAI procedures. This assessment gap highlights the necessity for additional XAI research and development to create reliable and widely recognized evaluation measures that can accurately gauge the interpretability and transparency of AI systems.

Moreover, an AI model's capability to predict new cases unseen in the training data is also challenging; several mechanisms can be used to achieve the robustness of an AI-based system. One example is data augmentation, transforming existing data, generating new data based on training data, or using transfer learning. Effective generalization of models

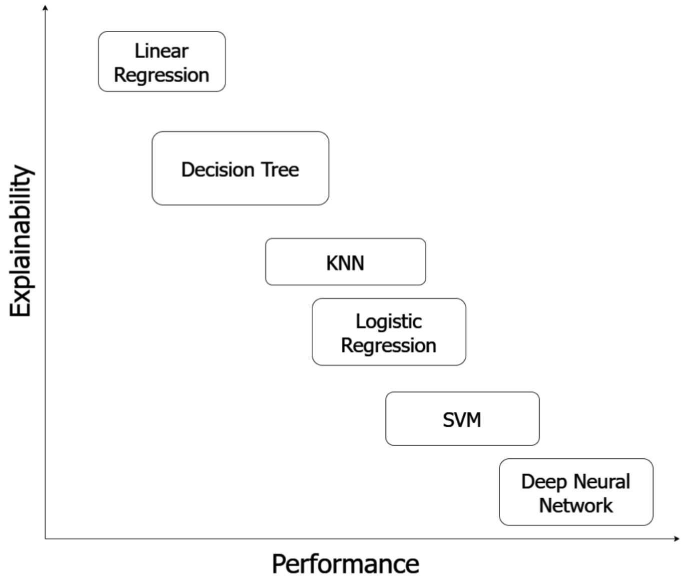

FIGURE 8. Performance-Interpretability trade-off: While the performance increases, the algorithm becomes more complex and hence less explainable.

is made possible through domain adaptation, which fills in the gaps between source and target domains. Methods such as instance-level and feature-level adaptability are essential. Concurrently, co-variate shift adapts to variations in input feature distributions, guaranteeing those models continue to function well in various scenarios [136]. By working together, these techniques improve AI systems' capacity to manage unknown data and unpredictable situations, opening the door to more robust and trustworthy AI.

In addition, adversarial defense techniques are utilized to safeguard machine learning models, particularly deep neural networks, against adversarial attacks. Adversarial defense seeks to enhance the resilience of AI classifiers to attacks by using permutation invariance. Unlike prior approaches, adversarial defense upholds the original accuracy of the classifiers, representing notable progress in AI safety [137].

Moreover, to guarantee data security and privacy in AI systems, the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) give recommendations about privacy rights and regulatory frameworks [138], [139]. While the GDPR provides ownership over the acquisition and use of personal data, the CCPA gives customers more control over their personally identifiable information. The confidentiality of private and sensitive information provided by data and models that may be shared throughout the AI system must be protected to create trustworthy AI systems.

Privacy and trustworthiness are closely linked, with AI systems being particularly threatening due to their inherent right to privacy. Effective data protection is crucial for mitigating security loss, including maintaining data validity, implementing access protocols, and interpreting data in a way that protects privacy. Distributed learning systems are particularly vulnerable to privacy breaches and data leaks. This is a significant issue in the case of medical data, where patient information can be leaked [140]. Several techniques have been used to ensure users' personal data protection from malicious activity and harm, such as Homomorphic encryption, differential privacy, and federated learning [141]. Federated learning is a new promising learning approach where clients and data owners collaborate on training a machine learning model without sharing their private data [142].

Finally, it's important to emphasize that even if AI systems are growing more autonomous, they must be kept under human oversight and control to ensure they work as designed and don't harm. They can be monitored and modified as necessary, as they should function within the constraints that their human founders established. However, protecting these systems becomes crucial as they become more powerful and sophisticated. Strong safety precautions must be put in place, such as fail-safe systems and moral standards, to avoid abuse and unintended consequences.

We can summarize the following findings regarding the research questions:

1) *RQ1:* We found that the safety of AI systems is directly related to how AI models learn; we discussed the safety of different learning techniques.
2) *RQ2:* We found that the researchers are interested in three main concerns: the verification and validation methods, the failure modes, and the autonomy, interaction, and control of AI systems.
3) *RQ3:* We found that the main challenges regarding AI safety are related to explainability and interpretability, robustness and reliability, fairness, and adversarial attacks.

VII. CONCLUSION AND FUTURE WORK

After examining the systematic review of AI safety, it's evident that the focus is on developing AI models that can be trusted. Recognizing the importance of incorporating safety considerations throughout the process, including data management, model development, and deployment, is becoming more widespread. Key areas of study in AI safety, such as learning methods, validation and verification techniques, failure modes, and managing AI autonomy, have been identified.

The main worries relating to safe AI, such as explainability, interpretability, robustness, reliability, fairness, bias, and adversarial attacks, have been thoroughly covered. Nevertheless, finding the right balance between performance and interpretability remains a significant challenge. As AI models become more effective, they also become more complex and less understandable.

Furthermore, the review highlights the need for additional research in defending against adversarial attacks as an essential aspect of designing secure models. The progression of research in safe AI underscores the significance of this area within the broader scope of AI advancement.

When we contemplate the concept of AI and autonomy, the interaction between humans and AI emerges as a pivotal consideration. The advancement of autonomous systems presents fresh challenges relating to safety and ethical concerns. It is imperative to guarantee that these systems can securely engage with humans and make decisions that align with human values and standards. This encompasses comprehending and appropriately responding to human inputs, articulating their actions in a manner understandable to humans, and assimilating feedback from humans.

In conclusion, AI safety is a crucial aspect of AI research that demands ongoing exploration and advancement.

It is hoped that this work will inspire further research and innovation in AI safety, leading to the development of more reliable, robust, and trustworthy AI systems that can interact effectively and safely with humans. This is particularly important as we move towards more Artificial General Intelligence (AGI) and fully autonomous AI systems.

A different approach might evaluate abstract problems like considering systems that might evade human control, outsmart humans, or also the ethical considerations such as trolley dilemma [143], or using GAN techniques such as deep fake to spread false information. Still, it was out of this study's scope, but it is suggested that future researchers may tackle this dilemma.

The findings discussed in this paper show the results of our exploratory study. The following phase is to build a safety framework with a set of recommended procedures for each of its processes and a thorough workflow to guide its implementation, helping to fill the gaps in the guidelines for future work. Our goal is to be able to apply a new method to safety-critical systems.

Finally, in future work, it's crucial to ensure that AI is created and applied in a morally and legally acceptable way. Continuous investigation and discussion of AI ethics are required to manage the difficulties presented by AI in operating autonomously and safely while maintaining performance.

REFERENCES

[1] M. Bearman and R. Luckin, "Preparing university assessment for a world with AI: Tasks for human intelligence," in *Re-Imagining University Assessment in a Digital World*. Cham, Switzerland: Springer, 2020, pp. 49–63.

[2] Y. Pan and L. Zhang, "Roles of artificial intelligence in construction engineering and management: A critical review and future trends," *Autom. Construction*, vol. 122, Feb. 2021, Art. no. 103517.

[3] D. L. Poole, A. K. Mackworth, and R. Goebel, *Computational Intelligence—A Logical Approach*. London, U.K.: Oxford Univ. Press, 1998.

[4] S. Russell and P. Norvig, *Intelligence Artificielle: Avec Plus de 500 Exercices*. Paris, France: Pearson Education, 2010.

[5] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, Mar. 2011.

[6] R. Sheh, "Explainable artificial intelligence requirements for safe, intelligent robots," in *Proc. IEEE Int. Conf. Intell. Saf. for Robot. (ISR)*, Mar. 2021, pp. 382–387.

[7] M. M. Hasan, M. U. Islam, and M. J. Sadeq, "Towards the technological adaptation of advanced farming through artificial intelligence, the Internet of Things, and robotics: A comprehensive overview," in *Artificial Intelligence and Smart Agriculture Technology*. New York, NY, USA: Auerbach, 2022, pp. 21–42.

[8] K. Matteucci, S. Avin, F. Barez, and S. Ó hÉigeartaigh, "AI systems of concern," 2023, *arXiv:2310.05876*.

[9] F. E. Morgan, B. Boudreaux, A. J. Lohn, M. Ashby, C. Curriden, K. Klima, and D. Grossman, "Military applications of artificial intelligence," RAND Corp., Santa Monica, CA, USA, Tech. Rep., 2020.

[10] J. C. Gore, "Artificial intelligence in medical imaging," *Magn. Reson. Imag.*, vol. 68, pp. A1–A4, May 2020.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, Jan. 2012, pp. 214–226.

[12] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[13] X. Zhao, A. Banks, J. Sharp, V. Robu, D. Flynn, M. Fisher, and X. Huang, "A safety framework for critical systems utilising deep neural networks," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12234, 2020, pp. 244–259.

[14] G. Bravo-Rocca, P. Liu, J. Guitart, A. Dholakia, D. Ellison, and M. Hodak, "Human-in-the-loop online multi-agent approach to increase trustworthiness in ML models through trust scores and data augmentation," in *Proc. IEEE 46th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jun. 2022, pp. 32–37.

[15] G. Rossolini, A. Biondi, and G. Buttazzo, "Increasing the confidence of deep neural networks by coverage analysis," *IEEE Trans. Softw. Eng.*, vol. 49, no. 2, pp. 802–815, Feb. 2023.

[16] H. He, J. Gray, A. Cangelosi, Q. Meng, T. M. McGinnity, and J. Mehnen, "The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 4, pp. 1398–1412, Dec. 2022.

[17] H. He, J. Gray, A. Cangelosi, Q. Meng, T. McGinnity, and J. Mehnen, "The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems," in *Proc. 3rd Int. Conf. Intell. Robotic Control Eng. (IRCE)*, Aug. 2020, pp. 68–74.

[18] P. Schmidt, F. Biessmann, and T. Teubner, "Transparency and trust in artificial intelligence systems," *J. Decis. Syst.*, vol. 29, no. 4, pp. 260–278, Oct. 2020.

[19] I. El Naqa and M. J. Murphy, *What is Machine Learning?*. Cham, Switzerland: Springer, 2015.

[20] C. Mujeeb Ahmed, M. A. Umer, B. S. S. Binte Liyakkathali, M. T. Jilani, and J. Zhou, "Machine learning for cps security: Applications, challenges and recommendations," in *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*. Cham, Switzerland: Springer, 2021, pp. 397–421.

[21] B. Paden, M. Cáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 33–55, Mar. 2016.

[22] A. Pereira and C. Thomas, "Challenges of machine learning applied to safety-critical cyber-physical systems," *Mach. Learn. Knowl. Extraction*, vol. 2, no. 4, pp. 579–602, Nov. 2020.

[23] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 1–39, May 2021.

[24] N. K. Gorelits, D. S. Kildishev, and A. V. Khoroshilov, "Requirements management for safety-critical systems. Overview of solutions," *Proc. Inst. Syst. Program. RAS*, vol. 31, no. 1, pp. 25–48, 2019.

[25] Z. Iqbal, "Assurance of machine learning/TinyML in safety-critical domains," in *Proc. IEEE Symp. Vis. Lang. Human-Centric Comput. (VL/HCC)*, Sep. 2022, pp. 1–2.

[26] H. G. Gurbuz, B. Tekinerdogan, C. Catal, and N. P. Er, "Test suite assessment of safety-critical systems using safety tactics and fault-based mutation testing," *Cluster Comput.*, vol. 27, no. 4, pp. 5377–5401, Jul. 2024.

[27] J. An, A. Mikhaylov, and K. Kim, "Machine learning approach in heterogeneous group of algorithms for transport safety-critical system," *Appl. Sci.*, vol. 10, no. 8, p. 2670, Apr. 2020.

[28] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: A comprehensive survey," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 945–990, 2022.

[29] R. Dhaya, R. Kanthavel, F. Algarni, P. Jayarajan, and A. Mahor, "Reinforcement learning concepts ministering smart city applications using IoT," in *Internet of Things in Smart Technologies for Sustainable Urban Development*. Cham, Switzerland: Springer, 2020, pp. 19–41.

[30] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 6292–6299.

[31] P. Ladosz, L. Weng, M. Kim, and H. Oh, "Exploration in deep reinforcement learning: A survey," *Inf. Fusion*, vol. 85, pp. 1–22, Sep. 2022.

[32] N.-M. Aliman and L. Kester, "Malicious design in AIVR, falsehood and cybersecurity-oriented immersive defenses," in *Proc. IEEE Int. Conf. Artif. Intell. Virtual Reality (AIVR)*, Dec. 2020, pp. 130–137.

[33] T. G. Rudner and H. Toner, "Key concepts in AI safety: An overview," CSET Issue Briefs, Pixley, CA, USA, Tech. Rep., 2021.

[34] S. D. Gribble, "Robustness in complex systems," in *Proc. 8th Workshop Hot Topics Operating Syst.*, May 2001, pp. 21–26.

[35] N. Kshetry and L. R. Varshney, "Safety in the face of unknown unknowns: Algorithm fusion in data-driven engineering systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8162–8166.

[36] J. Girard-Satabin, M. Alberti, F. Bobot, Z. Chihani, and A. Lemesle, "CAISAR: A platform for characterizing artificial intelligence safety and robustness," in *Proc. AISafety*, vol. 3215, 2022, pp. 1–9.

[37] S.-H. Choi, J.-M. Shin, P. Liu, and Y.-H. Choi, "ARGAN: Adversarially robust generative adversarial networks for deep neural networks against adversarial examples," *IEEE Access*, vol. 10, pp. 33602–33615, 2022.

[38] B. Tarchoun, A. B. Khalifa, and M. A. Mahjoub, "Investigating the robustness of multi-view detection to current adversarial patch threats," in *Proc. 6th Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2022, pp. 1–6.

[39] M. N. Akram, A. Ambekar, I. Sorokos, K. Aslansefat, and D. Schneider, "StaDRe and StaDRo: Reliability and robustness estimation of ML-based forecasting using statistical distance measures," in *Proc. Comput. Saf., Rel., Secur. (SAFECOMP)*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13415, 2022, pp. 289–301.

[40] D. Schwartz, Y. Alparslan, and E. Kim, "Regularization and sparsity for adversarial robustness and stable attribution," in *Advances in Visual Computing* (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)), vol. 12509. Cham, Switzerland: Springer, 2020, pp. 3–14.

[41] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," 2017, *arXiv:1702.02284*.

[42] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," 2021, *arXiv:2110.11334*.

[43] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Proc. 8th CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Dunhuang, China. Cham, Switzerland: Springer, 2019, pp. 563–574.

[44] J. Cooper, O. Arandjelović, and D. J. Harrison, "Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108743.

[45] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transp. Res. A, Policy Pract.*, vol. 94, pp. 182–193, Dec. 2016.

[46] X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F.-Y. Wang, "From features engineering to scenarios engineering for trustworthy AI: I&I, C&C, and V&V," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 18–26, Jul. 2022.

[47] A. Morales-Forero, S. Bassetto, and E. Coatanea, "Toward safe AI," *AI Soc.*, vol. 38, no. 2, pp. 685–696, Apr. 2023.

[48] A. Steimers and T. Bömer, "Sources of risk and design principles of trustworthy artificial intelligence," in *Proc. Int. Conf. Human-Comput. Interact.* Cham, Switzerland: Springer, 2021, pp. 239–251.

[49] J. Burden and J. Hernández-Orallo, "Exploring AI safety in degrees: Generality, capability and control," in *Proc. Workshop Artif. Intell. Saf. (SafeAI) 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 36–40.

[50] P. Festor, I. Habli, Y. Jia, A. Gordon, A. A. Faisal, and M. Komorowski, "Levels of autonomy and safety assurance for ai-based clinical decision systems," in *Proc. Comput. Saf., Rel., Secur. (SAFECOMP)*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12853, 2021, pp. 291–296.

[51] J. Hernández-Orallo, "AI safety landscape from short-term specific system engineering to long-term artificial general intelligence," in *Proc. 50th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Jun. 2020, pp. 72–73.

[52] P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Berlin, Germany: Springer, 2008, pp. 21–49.

[53] S. Dridi, "Supervised learning—A systematic literature review," *Preprint*, pp. 1–22, Dec. 2021.

[54] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *Proc. 3rd Int. Conf. Comput. for Sustain. Global Develop. (INDIACom)*, Mar. 2016, pp. 1310–1315.

[55] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Unsupervised learning," in *An Introduction to Statistical Learning: With Applications in Python*. Cham, Switzerland: Springer, 2023, pp. 503–556.

[56] L. Rokach and O. Maimon, "Clustering methods," in *Data Mining and Knowledge Discovery Handbook*. Boston, MA, USA: Springer, 2005, pp. 321–352.

[57] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, May 1996.

[58] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theory and applications," 2022, *arXiv:2205.10330*.

[59] F. Tambon, G. Laberge, L. An, A. Nikanjam, P. S. N. Mindom, Y. Pequignot, F. Khomh, G. Antoniol, E. Merlo, and F. Laviolette, "How to certify machine learning based safety-critical systems? A systematic literature review," *Automated Softw. Eng.*, vol. 29, no. 2, p. 38, Nov. 2022.

[60] L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, and J. K. Nurminen, "Systematic literature review of validation methods for AI systems," *J. Syst. Softw.*, vol. 181, Nov. 2021, Art. no. 111050.

[61] G. Vidot, C. Gabreau, I. Ober, and I. Ober, "Certification of embedded systems based on machine learning: A survey," 2021, *arXiv:2106.07221*.

[62] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert, "Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-making," *IEEE Comput. Intell. Mag.*, vol. 17, no. 1, pp. 72–85, Feb. 2022.

[63] A. S. Albahri, A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, J. Santamaría, C. Ouyang, A. Gupta, Y. Gu, and M. Deveci, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, pp. 156–191, Aug. 2023.

[64] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[65] Y. Wang and S. H. Chung, "Artificial intelligence in safety-critical systems: A systematic review," *Ind. Manage. Data Syst.*, vol. 122, no. 2, pp. 442–470, Feb. 2022.

[66] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Inf. Softw. Technol.*, vol. 64, pp. 1–18, Aug. 2015.

[67] *PRISMA*. Accessed: Apr. 4, 2023. [Online]. Available: http://prisma-statement.org/prismastatement/checklist.aspx

[68] M. J. Page et al., "PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, p. 160, Mar. 2021.

[69] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, and D. Moher, "Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement," *J. Clin. Epidemiol.*, vol. 134, pp. 103–112, Jun. 2021.

[70] W. M. Bramer, M. L. Rethlefsen, J. Kleijnen, and O. H. Franco, "Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study," *Systematic Rev.*, vol. 6, no. 1, pp. 1–12, Dec. 2017.

[71] L. Z. Atkinson and A. Cipriani, "How to carry out a literature search for a systematic review: A practical guide," *BJPsych Adv.*, vol. 24, no. 2, pp. 74–82, Mar. 2018.

[72] D. Manheim, "Multiparty dynamics and failure modes for machine learning and artificial intelligence," *Big Data Cogn. Comput.*, vol. 3, no. 2, p. 21, Apr. 2019.

[73] M. Giacobbe, M. Hasanbeig, D. Kroening, and H. Wijk, "Shielding Atari games with bounded prescience," in *Proc. AAMAS*, vol. 3, 2021, pp. 1495–1497.

[74] A. L. Buczak, B. D. Baugher, A. J. Berlier, K. E. Scharfstein, and C. S. Martin, "Explainable forecasts of disruptive events using recurrent neural networks," in *Proc. IEEE Int. Conf. Assured Autonomy (ICAA)*, Mar. 2022, pp. 64–73.

[75] F. Ritz, T. Phan, R. Müller, T. Gabor, A. Sedlmeier, M. Zeller, J. Wieghardt, R. Schmid, H. Sauer, C. Klein, and C. Linnhoff-Popien, "Specification aware multi-agent reinforcement learning," in *Agents and Artificial Intelligence* (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)), vol. 13251. Cham, Switzerland: Springer, 2022, pp. 3–21.

[76] P. T. Rajendran, H. Espinoza, A. Delaborde, and C. Mraidha, "Human-in-the-loop learning methods toward safe dl-based autonomous systems: A review," in *Proc. Comput. Saf., Rel., Secur. (SAFECOMP)*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12853, 2021, pp. 251–264.

[77] B. Adhikari, J. Peltomäki, S. B. Germi, E. Rahtu, and H. Huttunen, "Effect of label noise on robustness of deep neural network object detectors," in *Proc. Comput. Saf., Rel., Secur. (SAFECOMP)*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12853, 2021, pp. 239–250.

[78] L. B. Canonico and N. McNeese, "Flash crashes in multi-agent systems using minority games and reinforcement learning to test AI safety," in *Proc. Winter Simulation Conf. (WSC)*, Dec. 2019, pp. 193–204.

[79] J.-Y. Kim and S.-B. Cho, "Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck," in *Proc. CEUR-WS*, vol. 2560, 2020, pp. 105–112.

[80] X. Zhao, W. Huang, S. Schewe, Y. Dong, and X. Huang, "Detecting operational adversarial examples for reliable deep learning," in *Proc. 51st Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Supplemental Volume (DSN-S)*, Jun. 2021, pp. 5–6.

[81] A. Platzer, "The logical path to autonomous cyber-physical systems," in *Quantitative Evaluation of Systems* (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)), vol. 11785. Cham, Switzerland: Springer, 2019, pp. 25–33.

[82] N. Hunt, N. Fulton, S. Magliacane, T. N. Hoang, S. Das, and A. Solar-Lezama, "Verifiably safe exploration for end-to-end reinforcement learning," in *Proc. 24th Int. Conf. Hybrid Systems: Comput. Control*. New York, NY, USA: Association for Computing Machinery, May 2021.

[83] R. Kamoi and K. Kobayashi, "Out-of-distribution detection with likelihoods assigned by deep generative models using multimodal prior distributions," in *Proc. CEUR-WS*, vol. 2560, 2020, pp. 113–116.

[84] M. M. Sallami, M. I. Khedher, A. Trabelsi, S. Kerboua-Benlarbi, and D. Bettebghor, "Safety and robustness of deep neural networks object recognition under generic attacks," in *Neural Information Processing* (Communications in Computer and Information Science), vol. 1142. Cham, Switzerland: Springer, 2019, pp. 274–286.

[85] M. Maabreh, O. Darwish, O. Karajeh, and Y. Tashtoush, "On developing deep learning models with particle swarm optimization in the presence of poisoning attacks," in *Proc. Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2022, pp. 1–5.

[86] F. Arnez, H. Espinoza, A. Radermacher, and F. Terrier, "Improving robustness of deep neural networks for aerial navigation by incorporating input uncertainty," in *Proc. Comput. Saf., Rel., Secur. (SAFECOMP)*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12853, 2021, pp. 219–225.

[87] A. Krajna, M. Kovac, M. Brcic, and A. Šarcevic, "Explainable artificial intelligence: An updated perspective," in *Proc. 45th Jubilee Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, May 2022, pp. 859–864.

[88] K. Sokol and P. Flach, "Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety," in *Proc. CEUR-WS*, vol. 2301, 2019, pp. 1–4.

[89] M. Maabreh, A. Maabreh, B. Qolomany, and A. Al-Fuqaha, "The robustness of popular multiclass machine learning models against poisoning attacks: Lessons and insights," *Int. J. Distrib. Sensor Netw.*, vol. 18, no. 7, Jul. 2022, Art. no. 155013292211051.

[90] R. Srinivasan and A. Chander, "Understanding bias in datasets using topological data analysis," in *Proc. AISafety@ IJCAI*, vol. 2419, 2019, pp. 1–7.

[91] Y. Cai, "Safety analytics for AI systems," in *Proc. HCI Int. Late Breaking Papers, Multimodality Intell.*, vol. 12424, 2020, pp. 434–448.

[92] R. L. Castro and G. D. Rodosek, "Black box attacks using adversarial samples against machine learning malware classification to improve detection," 2018, pp. 16–20.

[93] V. Dementyeva, C. Hickert, N. Sarfaraz, S. Zanlongo, and T. Sookoor, "Runtime assurance for intelligent cyber-physical systems," in *Proc. ACM/IEEE 13th Int. Conf. Cyber-Physical Syst. (ICCPS)*, May 2022, pp. 288–289.

[94] P. Feifel, F. Bonarens, and F. Köster, "Reevaluating the safety impact of inherent interpretability on deep neural networks for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 29–37.

[95] N. Fulton and A. Platzer, "Safe AI for CPS," in *Proc. IEEE Int. Test Conf. (ITC)*, Oct. 2018, pp. 1–7.

[96] M. Ali, Y.-F. Hu, D. K. Luong, G. Oguntala, J.-P. Li, and K. Abdo, "Adversarial attacks on AI based intrusion detection system for heterogeneous wireless communications networks," in *Proc. AIAA/IEEE 39th Digit. Avionics Syst. Conf. (DASC)*, Oct. 2020, pp. 1–6.

[97] N. Jaipuria, K. Stevo, X. Zhang, M. L. Gaopande, I. C. Garcia, J. Jain, and V. N. Murali, "DeepPIC: Deep perceptual image clustering for identifying bias in vision datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4792–4801.

[98] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023.

[99] D. Ziegler, S. Nix, L. Chan, T. Bauman, P. Schmidt-Nielsen, T. Lin, A. Scherlis, N. Nabeshima, B. Weinstein-Raun, D. de Haas, B. Shlegeris, and N. Thomas, "Adversarial training for high-stakes reliability," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 9274–9286.

[100] J. Mattioli, H. Sohier, A. Delaborde, K. Amokrane-Ferka, A. Awadid, Z. Chihani, S. Khalfaoui, and G. Pedroza, "An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering," *AI Ethics*, vol. 4, no. 1, pp. 15–25, Feb. 2024.

[101] S. Mariani and F. Zambonelli, "Degrees of autonomy in coordinating collectives of self-driving vehicles," in *Proc. 9th Int. Symp. Leveraging Appl. Formal Methods, Verification Validation: Eng. Principles*, Rhodes, Greece. Cham, Switzerland: Springer, Oct. 2020, pp. 189–204.

[102] H. Wang, C. Li, J. Jiang, X. Zhang, Y. Zhao, and W. Gong, "Distribution-restrained softmax loss for the model robustness," 2023, *arXiv:2303.12363*.

[103] J. Martinez, A. Eguia, I. Urretavizcaya, E. Amparan, and P. L. Negro, "Fault tree analysis and failure modes and effects analysis for systems with artificial intelligence: A mapping study," in *Proc. 7th Int. Conf. Syst. Rel. Saf. (ICSRS)*, Nov. 2023, pp. 464–473.

[104] T. Hagendorff, "Linking human and machine behavior: A new approach to evaluate training data quality for beneficial machine learning," *Minds Mach.*, vol. 31, no. 4, pp. 563–593, Dec. 2021.

[105] B. Potteiger, T. Dignan, A. Mills, E. Pavelka, C. Frey, B. Nathan, M. Dagne, V. Garibaldi, and B. Otter, "Live virtual constructive environment for assuring the safety and security of complex autonomous vehicles," in *Proc. IEEE Int. Conf. Assured Autonomy (ICAA)*, Jun. 2023, pp. 53–56.

[106] B. D. Werner, B. J. Schumeg, T. M. Mills, and E. V. Velilla, "An assurance case for the DoD ethical principles of artificial intelligence," in *Proc. Annu. Rel. Maintainability Symp. (RAMS)*, Jan. 2023, pp. 1–7.

[107] T. Haider, K. Roscher, F. Schmoeller da Roza, and S. Günnemann, "Out-of-distribution detection for reinforcement learning agents with probabilistic dynamics models," in *Proc. Int. Conf. Auto. Agents Multiagent Syst.*, 2023, pp. 851–859.

[108] J. Pfrommer, M. Poyer, and S. Kiroriwal, "Reduce the handicap: Performance estimation for AI systems safety certification," in *Proc. IEEE 21st Int. Conf. Ind. Informat. (INDIN)*, Jul. 2023, pp. 1–7.

[109] A. Samadi, A. Shirian, K. Koufos, K. Debattista, and M. Dianati, "SAFE: Saliency-aware counterfactual explanations for DNN-based automated driving systems," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 5655–5662.

[110] Y. Wang, J. Xiao, Z. Wei, Y. Zheng, K.-T. Tang, and C. H. Chang, "Security and functional safety for AI in embedded automotive system—A tutorial," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 71, no. 3, pp. 1701–1707, Mar. 2024.

[111] M. Zeller, T. Waschulzik, R. Schmid, and C. Bahlmann, "Toward a safe MLOps process for the continuous development and safety assurance of ML-based systems in the railway domain," *AI Ethics*, vol. 4, no. 1, pp. 123–130, Feb. 2024.

[112] M. Fisher, V. Mascardi, K. Y. Rozier, B.-H. Schlingloff, M. Winikoff, and N. Yorke-Smith, "Towards a framework for certification of reliable autonomous systems," *Auto. Agents Multi-Agent Syst.*, vol. 35, no. 1, pp. 1–65, Apr. 2021.

[113] S. Mohseni, H. Wang, C. Xiao, Z. Yu, Z. Wang, and J. Yadawa, "Taxonomy of machine learning safety: A survey and primer," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–38, Aug. 2023.

[114] J. Yang, H. Wang, S. Wu, G. Chen, and J. Zhao, "Towards controlled data augmentations for active learning," in *Proc. ICML*, 2023, pp. 1–19.

[115] D. Gissin and S. Shalev-Shwartz, "Discriminative active learning," 2019, *arXiv:1907.06347*.

[116] M. He, Q. Tan, L. Cao, Q. He, and G. Jin, "Security enhanced optical encryption system by random phase key and permutation key," *Opt. Exp.*, vol. 17, no. 25, pp. 22462–22473, Dec. 2009. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?URI=oe-17-25-22462

[117] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, Mar. 2020.

[118] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[119] X. Gong, Q. Wang, Y. Chen, W. Yang, and X. Jiang, "Model extraction attacks and defenses on cloud-based machine learning models," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 83–89, Dec. 2020.

[120] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos, "Privacy and security issues in deep learning: A survey," *IEEE Access*, vol. 9, pp. 4566–4593, 2021.

[121] K. Tuyls and G. Weiss, "Multiagent learning: Basics, challenges, and prospects," *AI Mag.*, vol. 33, no. 3, pp. 41–52, Sep. 2012.

[122] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," 2016, *arXiv:1606.06565*.

[123] F. Cabitza and J.-D. Zeitoun, "The proof of the pudding: In praise of a culture of real-world validation for medical artificial intelligence," *Ann. Transl. Med.*, vol. 7, no. 8, p. 161, Apr. 2019.

[124] *ISO/CD 10303–226 Product Data Representation and Exchange. Application Protocol Part 226, Ship Mechanical Systems, N1015*, ISO Standard TC184/SC4/WG3, 2001.

[125] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.

[126] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, no. 6, pp. 601–611, 1976.

[127] R. Isermann, "Process fault detection based on modeling and estimation methods—A survey," *Automatica*, vol. 20, no. 4, pp. 387–404, Jul. 1984.

[128] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *Int. J. Human–Comput. Interact.*, vol. 36, no. 6, pp. 495–504, Apr. 2020.

[129] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Gener. Comput. Syst.*, vol. 135, pp. 364–381, Oct. 2022.

[130] B. Shneiderman, "Human-centered artificial intelligence: Three fresh ideas," *AIS Trans. Human-Comput. Interact.*, vol. 12, no. 3, pp. 109–124, 2020.

[131] J. Zhang and K. Cho, "Query-efficient imitation learning for end-to-end simulated driving," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 2891–2897.

[132] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, Dec. 2011, pp. 643–650.

[133] X. Zhao, K. Salako, L. Strigini, V. Robu, and D. Flynn, "Assessing safety-critical systems from operational testing: A study on autonomous vehicles," *Inf. Softw. Technol.*, vol. 128, Dec. 2020, Art. no. 106393.

[134] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[135] E. Dağlarli, "Explainable artificial intelligence (xAI) approaches and deep meta-learning models," in *Advances and Applications in Deep Learning*. London, U.K.: IntechOpen, 2020.

[136] T. Saikia, C. Schmid, and T. Brox, "Improving robustness against common corruptions with frequency biased models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11539–11551.

[137] J. Zhang, Y. Dong, M. Kuang, B. Liu, B. Ouyang, J. Zhu, H. Wang, and Y. Meng, "The art of defense: Letting networks fool the attacker," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3267–3276, 2023.

[138] *California Consumer Privacy Act (CCPA)*. Accessed: Mar. 28, 2024. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[139] *General Data Protection Regulation*. Accessed: Mar. 28, 2024. [Online]. Available: https://gdpr-info.eu/

[140] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, and R. Braren, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Mach. Intell.*, vol. 3, no. 6, pp. 473–484, May 2021.

[141] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[142] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 1354–1371.

[143] M. Cunneen, M. Mullins, F. Murphy, and S. Gaines, "Artificial driving intelligence and moral agency: Examining the decision ontology of unavoidable road traffic accidents through the prism of the trolley dilemma," *Appl. Artif. Intell.*, vol. 33, no. 3, pp. 267–293, Feb. 2019.

WISSAM SALHAB received the B.E. degree in computer and communication engineering from the Islamic University of Lebanon, Beirut, in 2019, and the M.S. degree in computer science from Lebanese University, Beirut, in 2022. He is currently pursuing the Ph.D. degree in computer science with the University of Québec at Chicoutimi, QC, Canada. His research interests include AI trustworthiness and safety in cyber-physical systems, human-AI interaction (HAII), responsible and robust AI, federated learning, and reinforcement learning.

DARINE AMEYED received the B.Sc. degree in computer science and management from the Institut Supérieure de Gestion, University of Tunis, in 2005, the M.Sc. degree in multimedia engineering from the University of Paris-Est Marne la-Vallée (University of Paris 11), France, in 2008, the M.Sc. degree in digital art and technology from the University of Rennes, Upper Brittany, France, in 2010, and the Ph.D. degree in engineering, software and information technologies from ÉTS, in 2017. Currently, she is an Adjunct Professor with Québec University at Chicoutimi and also the Head of IA with Quebec's Ministry of Transport. She was the Chief AI of Ministére de Cybetsècurité et Numérique in charge of Quebec's AI public sector strategy. Act as an Expert in United for Smart and Sustainable Cities (U4SSC)-International Telecommunication Union-ITU-United Nations. From 2015 to 2019, she was an Associate Researcher with ÉTS and the Scientific Program Manager of CIRODD, from 2015 to 2019. In addition, she has a multidisciplinary career in academia and the industrial sector, since 2005, holding positions of scientific and technology officers in several organizations, in information systems, artificial intelligence, and C-IoT platforms in Canada, Europe, and Africa. Her work focuses on ambient intelligence (AmI), the Cognitive-IoT, applied artificial intelligence, large-scale adoption of AI, privacy and security issues in AI, cyber-physical intelligent ecosystems, context-awareness, and human-centered computing. She received a certificate in applied AI (Columbia University, USA, in 2019).

FEHMI JAAFAR received the Ph.D. degree from the Department of Computer Science, Montreal University, Canada. He was a Researcher with the Computer Research Institute of Montreal, an Adjunct Professor with Concordia University, Edmonton, and a Postdoctoral Research Fellow with Queen's University and Polytechnique Montreal. He is currently an Associate Professor with Québec University at Chicoutimi and an Affiliate Professor with Laval University and Concordia University. His research has been published in top venues in computer sciences, including the *Journal of Empirical Software Engineering* (EMSE) and the *Journal of Software: Evolution and Process* (JSEP). He established externally funded research programs in collaboration with Defence Canada, Safety Canada, NSERC, and MITACS. His research interests include the Internet of Things security and the application of machine learning techniques in cybersecurity.

HAMID MCHEICK (Senior Member, IEEE) received the Ph.D. degree in computer science (software engineering and distributed systems) from the University of Montreal, Canada, in 2006. He is currently a Full Professor with the Computer Science Department, University of Québec at Chicoutimi, Canada. He has more than 25 years of experience in both academic and industrial areas. He is working on designing of smart and connected software applications, designing healthcare frameworks for many diseases, and designing smart Internet of Things and Cloud architectural models. He has supervised many post-doctorate, Ph.D., master's, and bachelor's students. He is an editor of a book with two researchers. He has also nine book chapters, more than 60 research papers in international journals, and more than 150 in international/national conferences and workshop proceedings in his credit. He has given many keynote speeches and tutorials in his research area, particularly in healthcare systems, pervasive and ubiquitous computing, distributed middleware architectures, software connectors, service oriented computing, the Internet of Things (IoT), mobile edge computing, fog computing, and cloud computing. He has received many grants from governments, industries, and academics. He is a chief editor, the chair, the co-chair, a reviewer, and a member of many organizations (such as ACM, Springer, Elsevier, Wiley, and Inderscience) around the world.