IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 November 2022, accepted 6 December 2022, date of publication 14 December 2022,
date of current version 20 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3229233

SURVEY

Safety Assurance of Artificial Intelligence-Based Systems: A Systematic Literature Review on the State of the Art and Guidelines for Future Work

ANTONIO V. SILVA NETO, JOÃO B. CAMARGO JR., JORGE R. ALMEIDA JR., AND PAULO S. CUGNASCA

Safety Analysis Group (GAS), Department of Computer Engineering and Digital Systems (PCS), Escola Politécnica, Universidade de São Paulo (USP), São Paulo 05508-010, Brazil

Corresponding author: Paulo S. Cugnasca (cugnasca@usp.br)

The work of Antonio V. Silva Neto was supported in part by the Brazilian Institution CAPES—Coordenação de Aperfeiçoamento de Pessoal de Nível Superior through PROEX Scholarship under Grant 88887.513631/2020-00, and in part by the Brazilian institution FDTE—Fundação para o Desenvolvimento Tecnológico da Engenharia through Scholarship under Grant 1954.01.20.

ABSTRACT The objective of this research is to present the state of the art of the safety assurance of Artificial Intelligence (AI)-based systems and guidelines on future correlated work. For this purpose, a Systematic Literature Review comprising 5090 peer-reviewed references relating safety to AI has been carried out, with focus on a 329-reference subset in which the safety assurance of AI-based systems is directly conveyed. From 2016 onwards, the safety assurance of AI-based systems has experienced significant effervescence and leaned towards five main approaches: performing black-box testing, using safety envelopes, designing fail-safe AI, combining white-box analyses with explainable AI, and establishing a safety assurance process throughout systems' lifecycles. Each of these approaches has been discussed in this paper, along with their features, pros and cons. Finally, guidelines for future research topics have also been presented. They result from an analysis based on both the cross-fertilization among the reviewed references and the authors' experience with safety and AI. Among 15 research themes, these guidelines reinforce the need for deepening guidelines for the safety assurance of AI-based systems by, e.g., analyzing datasets from a safety perspective, designing explainable AI, setting and justifying AI hyperparameters, and assuring the safety of hardware-implemented AI-based systems.

INDEX TERMS Artificial intelligence, formal verification, learning systems, machine learning, neural networks, product safety engineering, risk analysis, safety.

I. INTRODUCTION

With the Fourth Industrial Revolution (i.e., Industry 4.0), the conception and practical usage of Intelligent Cyber-Physical Systems (ICPS) relying on Artificial Intelligence (AI) to perform safety-critical functions is likely to increase in the following years, leveraged especially by the vigorous ongoing research on the matter for autonomous ground vehicles [1], [2] which represent a major paradigm shift from current transportation applications [3].

ICPSs are highly likely to be introduced even in application areas with strict safety-related requirements and which have been historically refractory to significant design paradigm shifts, such as the aeronautics and railway domains. A roadmap of future objectives published by the European Union Aviation Safety Agency (EASA) in 2020, for instance, aims that the first safety-critical AI-based systems shall be certified up to 2025 and that such safety certification shall be extended to fully autonomous systems up to 2035 [4]. In the railway domain, the European Union Agency for Railways (ERA) executive director claims that AI-based safety-critical systems are necessary in the forthcoming years to keep rail transportation means competitive against other means

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

of transportation and foresees major architecture changes on future generation AI-based systems, which may be fully distributed aboard trains instead of featuring centralized elements as in current Communication-Based Train Control (CBTC) systems [5].

Based on this context, the safety assurance of AI-based systems is deemed of paramount importance to allow the successful deployment of these systems in their respective applications [6]. Safety assurance is herein defined as the set of activities, means, and methods that shall be considered, throughout the lifecycle of a system, to produce results towards building arguments that confidently support the safety requirements / targets of such a system have been met. This concept extends the definitions of 'safety assurance' coined by McDermid et al. [7] and Habli et al. [8] with the 'safety management' concept of several standards of the safety métier, such as IEC61508:2010, DO-254:2000, CENELEC EN50126-1:2017, and CENELEC EN50129: 2018, related to the process of building and maintaining safety arguments throughout the lifecycle of a system [9], [10], [11], [12]. The original definition of 'safety assurance' presented by McDermid et al. is "(...) *justified confidence or certainty in a system's capabilities, including its safety*", whereas Habli et al. [8], state that "*safety assurance is concerned with demonstrating confidence in a system's safety*".

The first step towards improving the safety assurance of AI-based systems is establishing the state of the art of scientific and technical advance in the theme and identifying potential gaps for future research. Even though reviews with similar motivation which have been identified, they are not deemed able to fully characterize such a theme due to the following reasons, further detailed and justified within this paper's section II:

a) Given the significant effervescence of research on means, methods and tools to support the design, verify, and validate safety-critical AI from 2018 onwards, there is an increasing need to keep track of these updates and report them in a didactic yet detailed way to the research community. Hence, even rather recent literature reviews, published in the last couple of years, are potentially unable to fully capture the current landscape towards assuring the safety of AI-based systems;
b) There are literature reviews focused on exploring the safety assurance of AI-based systems in specific application domains only (e.g., autonomous vehicles). As a result, it is deemed that they are potentially restrained in reporting relevant general-purpose research for safety-critical AI-based systems as a whole;
c) Guidance for future research on the safety assurance of AI-based systems has not been detected on some literature reviews. As a result, one might not be able to easily identify themes that might still be of interest for the research community;
d) Finally, there are reviews which lack proper methodological systematization or whose authors themselves suggest that additional investigation of the literature shall be carried out in the future.

Hence, this research has been idealized to fill the aforementioned gaps of preexisting literature reviews on the theme. Hence, the objectives of this paper are to (i.) present the state of the art the safety assurance of AI-based systems, including methods and techniques to do so, and (ii.) identify the main challenges needing further research – notably towards a general-purpose, application-independent method with guidelines for the safety assurance of AI-based systems. For that purpose, a Systematic Literature Review (SLR) of peer-reviewed material formally published up to August 26th, 2022 has been carried out along with critical cross-fertilization among the reviewed research to draw more extensive conclusions on both objectives.

The remainder of the paper is structured in six sections. Section II aims to justify the contribution of this research, notably comparing and contrasting it with other SLR-based papers. The SLR itself is covered in sections III to VI. The SLR method is presented in section IV, whereas the SLR results themselves are split into three parts: bibliometrics analyses are discussed in section V, the state of the art on the safety assurance of AI-based systems is presented in section VI, and the guidelines for future work in the area is covered in section VI. Finally, section VII closes the paper with the conclusions of the research.

II. CONTRIBUTION JUSTIFICATION: DIFFERENCES FROM OTHER LITERATURE REVIEWS

During this research, the literature review-oriented papers by Ballingall et al. [13], Chia et al. [14], Dey and Lee [15], Kabir [16], Nascimento et al. [17], Rajabli et al. [18], Rawson and Brito [19], Siedel et al [20], Tahir and Alexander [21], Tambon et al. [22], Wang and Chapman [23], Wang and Chung [24], Wen et al [25], Zhang and Li [26], and Zhang et al. [27] have been identified as somehow relating safety to AI-based systems. The objective of this section is to justify that the present SLR either differs from these or goes beyond their scope, hence supporting the contribution of the present research as a broader and deeper literature review on the safety assurance of AI-based systems.

The approach employed by Chia et al. [14], Nascimento et al. [17], Rajabli et al. [18], and Tahir and Alexander [21] focuses on safety analysis findings specifically related to autonomous ground vehicles (i.e., autonomous cars). Similarly, Rawson and Brito [19] have concentrated their research on reviewing the application of AI on maritime applications, notably on automating the prediction and the assessment or risks and accidents involving automated ships. The approach used in the present SLR not only includes both of the aforementioned applications, but it also goes beyond them in exploring research not only related to other application domains, but also unrelated to specific target applications (i.e., which concentrates on general-purpose technical aspects of AI and safety).

The review presented by Kabir [16] focuses on how to employ Fault Tree Analysis (FTA) and its extensions on Model-Based Dependability Analysis, hence not fully addressing the problem of AI-based systems safety assurance. In the present SLR, several safety analysis techniques in addition to FTA are also covered; moreover, the scope of the present SLR is broader, as approaches and gaps on the safety assurance of AI-based systems are also discussed.

The SLR performed by Zhang and Li [26] has the objective of analyzing methods and approaches for the testing and the verification of AI-based systems and identifying challenges and gaps for future studies on the area. Since testing and verification are means to build safety arguments to ensure that a safety-critical is safe, this work is deemed relevant and somewhat overlapping to this SLR in that regard. On the other hand, the SLR of Zhang and Li [26] has two limitations which have been overcome on the present SLR. Firstly, Zhang and Li [26] have restricted their analyses to neural networks only; secondly, the reviewed publications are within the timespan 2011-2018. On the herein reported SLR, the scope of AI approaches, techniques and algorithms has been significantly broadened– hence not restrained to neural networks -, and, ultimately, research published up to August 26th, 2022 has been considered. Based on the effervescence of research on AI-based safety-critical systems identified and discussed in this paper, significant advancements have occurred since 2018.

The SLR performed by Tambon et al. [22] shares similarity with the present research with regard to the research objectives themselves – namely, (i.) providing a landscape on means and methods for assuring that AI systems are sufficiently safe for their certification and (ii.) suggesting future work yet to be explored in the area. Despite these similarities, there are methodological and scope differences which support the relevance of the herein reported SLR in contributing with the safety assurance of AI-based systems.

Firstly, Tambon et al. [22] have focused their review efforts specifically on software-implemented machine learning, whereas only other types and implementations of AI are explored in the present SLR (for instance, knowledge-based systems, hardware-implemented AI). Secondly, the controlled search vocabulary utilized by Tambon et al. [22] is more restrictive than the one considered in the present research with regard not only to expressions for the safety and AI métiers, but also to potential application domains (only transportation on Tambon et al. [22], and unrestricted in the present SLR).

Finally, Tambon et al. [22] have constrained the themes of interest for future work to six major topics, therein referred to as 'robustness', 'uncertainty and out-of-distribution', 'explainability, 'formal and non-formal verification', 'safety considerations in reinforcement learning', and 'direct certification'. In the present SLR, these themes have been covered (albeit with potentially different names) along with others, making up for a total of 15 major research areas for future work towards the safety assurance of AI-based systems.

The literature reviews presented by Ballingall et al. [13], Dey and Lee [15], Siedel et al. [20], Wang and Chapman [23], Wang and Chung [24], Wen et al. [25], and Zhang et al. [27] all share two main similarities with the herein presented SLR: (i.) the lack of scope limiting to specific applications and (ii.) the objective of presenting an overview of safety analysis techniques for intelligent systems. However, all of them differ from the present one in aspects that make the latter broader and/or more accurate on depicting the state of the art and the gaps on the safety assurance of AI-based systems. This is justified for each of the six aforementioned reviews on the following paragraphs.

The literature review by Ballingall et al. [13] was not carried out systematically, and the authors not only justify studying the safety assurance of AI-based systems based on a single application (automated driving systems), but they also conclude that future investigation on the matter is still needed. Based on these remarks, the aim of the present research is to fill the gap of the review by Ballingall et al. [13] by means of a systematic and reproducible literature review spanning a broader search range.

The SLR performed by Dey and Lee [15] comprise three limitations, namely (i.) potentially non-peer-reviewed papers (e.g., available on arXiv), (ii.) conflicting information regarding the timespan of the considered papers (2005-2020 and 2015-2020 intervals are mentioned by the authors) and (iii.) brief discussion on future work, which are deemed better addressed in the present SLR. This is justified by four arguments: (i.) covering a wider range of official peer-reviewed reference databases (e.g., Engineering Village and Web of Science), (ii.) disregarding information from research which has not been formally published yet (e.g., arXiv-sourced papers have not been considered in this SLR), (iii.) defining a clear and wide timespan to the considered publications (all papers published until August 26th, 2022 with no starting date limit) and (iv.) dedicating a full paper section to the discussion of future work stemming from the gaps identified in current research on the safety assurance of AI-based systems.

Siedel et al. [20], in turn, have four main limitations. First, the controlled vocabulary used in searches comprises limited expressions from the safety and AI areas and also includes marginally-related topics, such as reliability. Secondly, Scopus was the only search engine considered by the authors. Thirdly, the keywords from the search vocabulary were checked only on the title of publications. Finally, the authors have not explored how the safety assurance of AI-based systems has evolved with time, nor captured technical trends of the area for future work. In the present SLR, an in-depth search language enriched in both wide and depth of expressions related to safety and AI has been crafted. Moreover, the search domain has been expanded to four other search engines other than Scopus and, in addition to the publication titles, searches have also been performed within the abstracts and the keywords of the indexed publications. Finally, a detailed landscape of the safety assurance of AI-based systems has also been presented. It includes bibliometrics, trends of the

area throughout its past and present, and guidelines for future work.

The SLR developed by Wang and Chapman [23] is limited to presenting the link between risk analyses and the control of autonomous systems, focusing on reviewing the main variants and algorithms of AI that are used on such applications and how their safety is ensured. In the present SLR, safety-critical AI applications other than the risk analysis of autonomous systems have also been covered, such as the usage of AI within the core control of safety-critical systems). Such an expansion of scope also allowed identifying more means to potentially build safety-critical AI-based systems and ensure their safety in comparison to those observed by Wang and Chapman [23].

Wang and Chung [24] have performed an SLR aiming to describe how AI has been used in safety-critical systems and propose potential future work to further promote such usage. Despite the partial convergence of results obtained by Wang and Chung [24] with the conclusions of the present study, there are noteworthy remarks that justify the relevance of the herein reported SLR. Firstly, the search expressions used by Wang and Chung [24] to characterize the AI and safety métiers are simpler and more restrictive than the ones of this SLR. Secondly, Wang and Chung [24] have added dependability within the scope of the SLR: since dependability comprises concepts other than safety, such as reliability and availability, some results obtained by the authors are not related to safety per se. In order to circumvent this, the present SLR has been conceived with a tighter link to the safety area. Thirdly, alike Dey and Lee [15], non-peer-reviewed papers (e.g., available on arXiv) have been considered by Wang and Chung [24], whereas only publications which have been formally published after acceptance on peer reviews have been taken into consideration in the present SLR. These limitations also translate into the numbers of retrieved and analyzed publications: while Wang and Chung [24] have assessed 3087 research papers and identified 92 of them as potentially relevant, the herein SLR starts with a set of 5090 publications, among which 329 were deemed relevant for the safety assurance of AI-based systems.

The SLR performed by Wen et al. [25] has as its main limitation the fact that the reviewed papers were randomly sampled from a set of papers. The authors themselves claim that a major contribution with their work is to exhaustively analyze publications on the safety assurance of AI-based systems, which is exactly one of the present research's objectives.

Finally, the SLR published by Zhang et al. [27] is the one that resembles the most the present SLR with regard to identifying relevant future work on the safety assurance of AI-based systems. Despite such similarities, Zhang et al. [27] lags behind the herein documented research on two main aspects: (i.) the criteria employed to collect and review reference studies, which is not clearly stated by the authors, and (ii.) the lack of a detailed description on how the state of the art of AI-based safety-critical systems has evolved with time up to the present time. In this SLR, a full section has been devoted to presenting the work method, and two sections, to characterizing how the relationship between AI and safety has emerged and progressed up to 2022. Furthermore, the guidelines for future work also include further themes and additional discussions on feasibility which remained unexplored by Zhang et al. [27].

III. SYSTEMATIC LITERATURE REVIEW METHOD

The objective of this section is to present the SLR process which bases the findings and the conclusions of this research. This section has been structured in such a way to (i.) provide foundation that the literature review method is systematic and sound for the research purpose, and (ii.) define and present information and notation from the SLR itself which is utilized on the remainder of the paper.

The SLR was carried out with six main activities in the following order: (A) the definition of the search keywords, (B) the decision on the search engines which were part of the searches scope, (C) the collection of search results and duplicate removal, (D) the Title-Abstract-Keywords (TAK) filtering, (E) the definition of the questionnaire for the full-text semantic filtering, and, finally, (F) the full-text semantic filtering.

These activities are illustrated in the workflow of Figure 1 and detailed in the forthcoming subsections. Since the analyses corresponding to steps (C), (D), and (F) are progressively finer filters towards obtaining relevant research papers on the safety assurance of AI-based systems, the SLR process depicted in Figure 1 was shaped as an horizontal funnel-shaped process with three filtering stages, each of which representing one of the aforementioned steps. Moreover, the numeric results depicted in Figure 1 for steps (C), (D) and (F) will be explored in section IV.

Similarly to Nascimento et al. [17], the tasks suggested by Asadollah et al. [28] and Petersen et al. [29] were employed to guide the crafting of keywords considered in step (A), as well as the questionnaire of step (E). Furthermore, all activities of the SLR were led by the first author or this paper (A. V. Silva Neto), and the results were discussed with other researchers in walkthroughs in order to ensure their validity and the adjudication of potential conflicts.

A. DEFINITION OF SEARCH KEYWORDS SLR SEARCH LANGUAGE

In order to formally define logical expressions to guide the searches of relevant publications on the search engines, a formal regular search language, herein called 'SLR Search Language', was conceived using Wirth's notation. An overview of the structure of the SLR Search Language is presented in Figure 2, in which 'AND', 'OR', and 'NOT' gates are employed to express the relationship among the SLR Search Language expression groups.

The SLR Search Language comprises expressions which shall simultaneously satisfy three criteria: (i.) the presence of terms related to the safety assurance domain ('Safety Assurance Area' on Figure 2), (ii.) the presence of terms

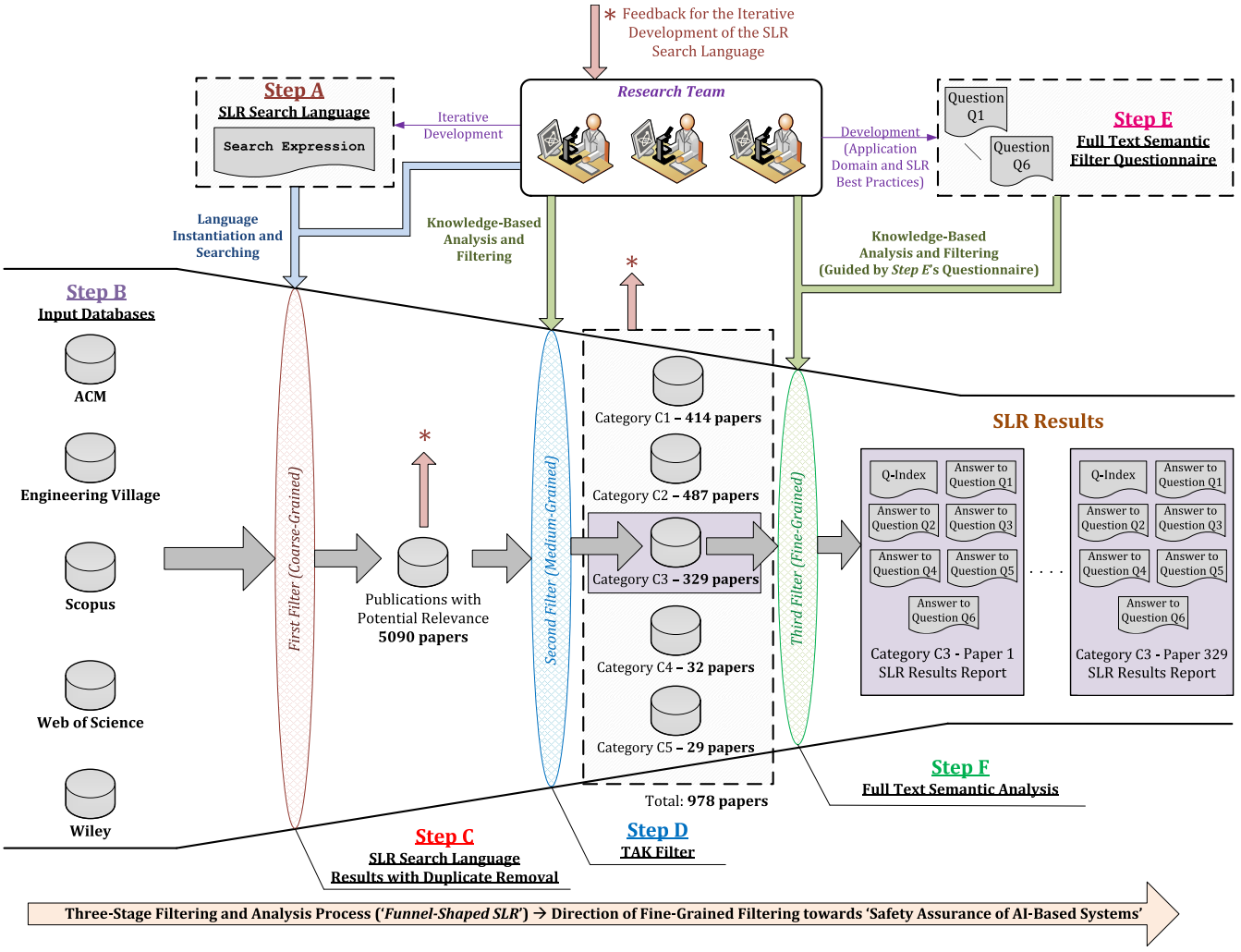

FIGURE 1. SLR process overview with numeric results of steps C, D and F.

related to AI ('AI Area' on Figure 2), and (iii.) the absence of terms related to applications which are not related to the technical safety domain ('Unrelated Area' on Figure 2). The SLR Search Language was iteratively crafted in preliminary phases of the SLR based on two aspects: (i.) the relevance of expressions for the scope of the research, and (ii.) the retrieval of a quantity of publications feasible for the objectives of this SLR whilst maximizing as much as possible relevant papers on the safety assurance of AI-based systems.

The expressions of the 'Safety Assurance Area' shall include at least one term from the 'Safety Assurance Process Items' subgroup and one term from the 'Safety Assurance Systematization' subgroup, as per Figure 2. The 'Safety Assurance Process Items' subgroup includes a set of verbal and non-verbal expressions (on singular and plural forms whenever applicable) with relevant activities and/or products of the safety assurance process. The safety assurance activities that have been considered are 'safety analysis', 'safety assessment', 'safety assurance', 'safety verification', 'safety validation', 'safety evaluation', 'risk analysis', 'risk assessment', 'risk evaluation', 'hazard analysis', 'hazard assessment', and 'hazard evaluation'. The safety assurance products, in turn, include 'safety case', 'assurance case', and 'assurance pattern'. Even though 'assurance pattern' is not a safety assurance result per se, it was considered part of the SLR Search Language given the increasing efforts on building reusable *de facto patterns* for assuring the safety of similar AI-based systems [30].

The 'Safety Assurance Systematization' subgroup, in turn, comprises expressions that are related to systematizing a systems safety lifecycle. It includes the words 'method', 'methodology', 'technique', 'approach', 'framework', and their corresponding plurals.

The 'AI Area' group reunites terms related to the artificial intelligence métier. These are divided in six subgroups: 'General AI', 'Machine Learning', 'Adaptive Systems', 'Industry 4.0', 'Data Science', and 'Knowledge-Based Systems' (formerly known as 'expert systems'). The 'OR' gate interconnecting them in Figure 2 indicates that each terminal

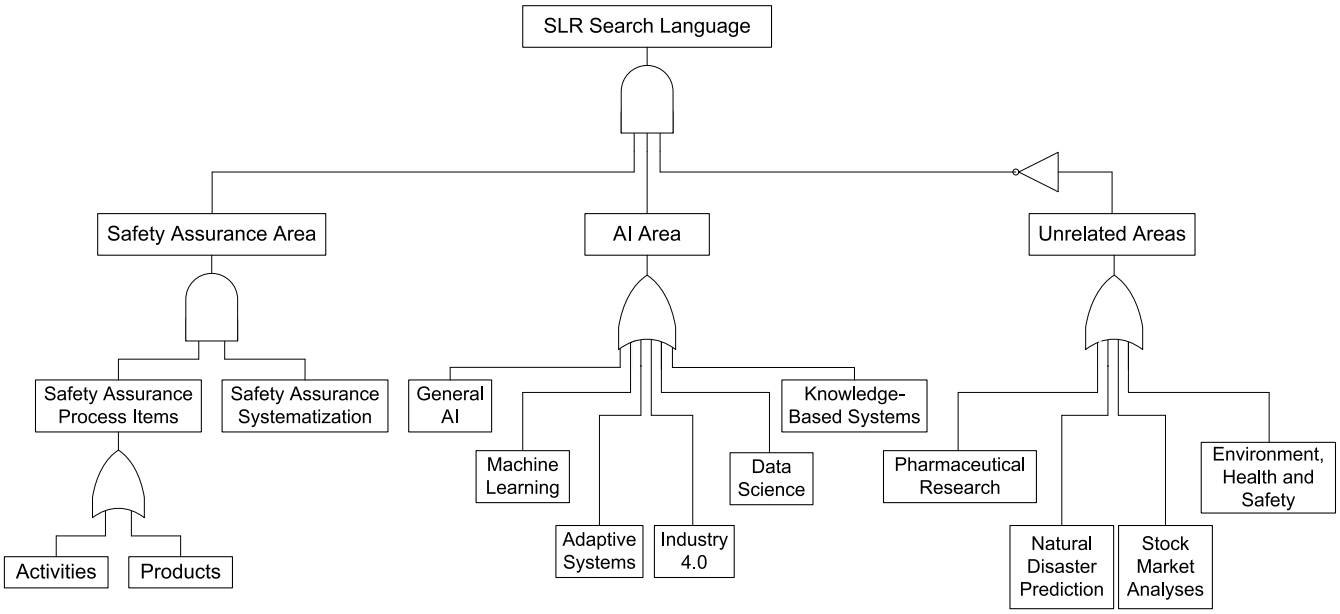

FIGURE 2. Overview of the SLR search language logical structure.

expression from any group suffices as a search expression for 'AI Area'.

Each subgroup's keywords have been defined to cover the high level definitions of each subgroup (e.g., 'artificial intelligence', 'AI', 'machine learning', 'data mining', etc.). Wherever applicable, two other subsets of expressions have also been included within each subgroup: (i.) the names of relevant AI variants (e.g., 'supervised learning', 'reinforcement learning', 'data clustering'), and (ii.) relevant formalisms and algorithms that implement the corresponding AI variants (e.g., 'neural network').

Finally, the 'Unrelated Areas' group has been iteratively built as a non-exhaustive set of expressions which shall lead to the exclusion of texts with them regardless of expressions from 'Safety Assurance Area' and 'AI Area' groups being satisfied. The 'Unrelated Areas' group expressions include four main application domains that have been manually identified as unrelated to the safety assurance of AI-based systems at the initial stages of the SLR – namely, pharmaceutical research (drug(s)), natural disaster prediction (flood(s), earthquake(s)), stock market analyses (credit(s), asset(s), insurance(s), portfolio(s), investment(s), stock(s)), and environment, health and safety (EHS) as a whole. The EHS field comprises public health policies (pregnancy, drug(s)), ergonomics, and issues related to occupational risk management of the oil industry, coal mines, dams, and construction sites.

B. CHOICE OF SEARCH ENGINES

With the exception of ScienceDirect and SpringerLink, all of the remaining search engines employed by Nascimento et al. [17] – namely ACM, Engineering Village, Scopus, Web of Science, and Wiley – were considered in this SLR. Pre-prints (e.g., sourced directly from private repositories on arXiv and Zenodo) are pruned at this step for quality concerns.

ScienceDirect was not considered on this SLR because its TAK-related indexed content, which is the starting point for this SLR, is entirely within Scopus [31]. SpringerLink, in turn, was disregarded after it has been assessed that relevant Springer-sourced publications, both periodic (e.g., journals) and non-periodic (e.g., conference proceedings and book chapters), have been successfully captured by means of Engineering Village, Scopus, and Web of Science.

C. COLLECTION OF SEARCH RESULTS AND DUPLICATE REMOVAL

The results of this research comprise references retrieved from TAK searches performed in all search engines from the previous section on August 26th, 2022. The results were exported in either BibTeX or Research Information Systems (RIS) formats and loaded onto Mendeley and JabRef tools for a semiautomatic duplicate removal (i.e., aided by tools but manually confirmed or rejected case by case).

D. TAK FILTER

The TAK filter was based on manual semantic analysis of TAK information of each and every reference retrieved from the previous step and allowed classifying the obtained results into six categories (C0 to C5). These categories were defined taking into account a didactic approach to split the obtained results into proper semantic groups related to this research up to some extent.

- **C0**: Not relevant to the research;
- **C1**: Research on AI applied in off-line safety assurance;
- **C2**: Research on AI applied in safety-critical functions, but without evidence of safety assessment per se;

- **C3**: Research on the safety assurance of AI-based safety-critical functions;
- **C4**: Contextualization of AI on Industry 4.0 applications;
- **C5**: Research on AI employed in security-critical functions with potential impact on safety.

It is worth noting that a reference can be classified into more than a single category from C1 to C5 based on its TAK information, since these categories are not mutually exclusive. Since the aim of the research is to study the safety assessment of AI-based safety-critical systems, category C3 is considered the most important for that purpose, whereas the other ones (except for C0) are deemed marginally relevant due to the following reasons:

a) Studies within categories C1 and C2 provide examples of AI-based safety-critical systems and applications which can, thus, benefit from the safety assurance of AI-based systems;
b) References within categories C4 and C5 deal with correlate themes and, as a result, shed some light on the contextualization of this research and may also provide guidance for potential future work.

E. QUESTIONNAIRE FOR THE FULL-TEXT SEMANTIC FILTER

Six questions (Q1 to Q6) were conceived to extract relevant information from the full-text review of references in line with the SLR objectives. The questionnaire has been only applied to category C3 references, since these are directly related to the research theme (i.e., safety assurance of AI-based systems).

- **Q1**: What is the objective of the research?
- **Q2**: Which AI techniques were considered in the research?
- **Q3**: How has the safety assurance of AI been considered within the study?
- **Q4**: Which results were obtained on the research (including potential shortcomings)?
- **Q5**: Which future research topics were identified by the authors?
- **Q6**: What other strengths and weaknesses were identified in the study during its review?

It is worth noting that Q6 resorts to the researchers' knowledge in assessing positive and negative aspects of the reviewed references. This is an important question to fulfill the objective of this research in providing an overview of future work which goes beyond what is directly proposed by the reviewed references' authors.

F. FULL-TEXT SEMANTIC FILTER

In this step, the full text of every reference on category C3 is reviewed based on the questionnaire defined in step E. A brief report of the results obtained for each text is developed, and the results of these reports are compiled based on the objectives of this SLR.

In addition to questions Q1 to Q6, an integer quality metric ranging from 0 to 6 – herein referred to as Q-index – has been crafted in order to rate how well each reference contributes to the objectives of the present research based on its overall quality and the covered topics. The Q-index includes two definitions: a discrete definition for each of the valid integer values within the interval [0; 6] and a categorized definition on three fuzzy groups: low quality, average quality and high quality.

Low Quality References

- $\mathbf{Q = 0}$: Highly restrictive relevance;
- $\mathbf{Q = 1}$: Weak relevance.

Average Quality References

- $\mathbf{Q = 2}$: Fair relevance;
- $\mathbf{Q = 3}$: Sufficient relevance.

High Quality References

- $\mathbf{Q = 4}$: Above average relevance;
- $\mathbf{Q = 5}$: Very good relevance;
- $\mathbf{Q = 6}$: Strong relevance.

IV. SYSTEMATIC LITERATURE REVIEW BIBLIOMETRICS RESULTS

The objective of this section is to present an overview of the bibliometrics extracted from the SLR results. Such an analysis is deemed relevant because it allows characterizing major trends on how research which joins AI and safety assurance has evolved with time. The first analysis, presented in subsection IV-A, is based on how the number of reviewed references of each category C1 to C5 has evolved with time.

Since the focus of this SLR is on the safety assurance of AI-based systems and such theme corresponds to the scope of C3-categorized publications, the bibliometrics analyses of C3 are enriched by correlating them to the Q-index attributed to each C3 publication as part of the SLR method. This analysis is covered in subsection IV-B.

Finally, the concluding remarks of the bibliometrics analysis and the justification of its importance to the remainder of the research are summarized in subsection IV-C.

A. OVERALL BIBLIOMETRICS FOR CATEGORIES C1 TO C5

A total of 5090 references, as shown in Figure 1, was obtained after filtering duplicates from the set of results retrieved on August 26th, 2022 from the search engines listed in subsection III-B when applying the SLR Search Language defined in subsection III-A.

After applying the TAK filter to the 5090 references obtained at the previous step, 4112 of them (80.8%) were included into category C0 and, hence, they were not considered relevant for this research. The remaining 978 (19.2%) were classified as part of at least one of the categories C1 to C5 defined in subsection III-D according to the quantities presented in Figure 1 for each category – namely, 414 references on C1, 487 references on C2, 329 references on C3, 32 references on C4 and 29 references on C5.

There are two reasons why summing the results for each of the C1 to C5 categories yields a result greater than the 978 references which were classified as somehow relevant

by means of the TAK filter. The first one, discussed in subsection III-D, is that categories C1 to C5 are not mutually exclusive. The second reason is that, since the TAK filter is somehow coarse to ensure proper classification of every single reference into their actual categories, some references were also conservatively classified in additional categories whenever the latter ones could not be categorically ruled out. This is especially important for C3, as the analysis of all references within it would mandatorily progress up to their full-text semantic analysis (as mentioned in subsections III-E and III-F).

The graph presented in Figure 3 depicts the evolution on the quantity of publications which have been classified as part of C1, C2, C3, C4 and C5 between 1986 and 2022 (up to August 26th). 1986 corresponds to the year of the very first reference relating AI to safety assurance.

Two main results can be inferred from Figure 3. The first result is that, after initial exploratory research carried out up to the mid-2000s, there have been two waves of significant increase in research correlating safety to AI. The first of them occurred between 2008 and 2014 and has been mostly concentrated on C1 and C2 publications, Hence, one can infer that the relationship between AI and safety on this first research wave has at most dealt with using AI as a tool to support the safety analysis of safety-critical systems, regardless of AI being present in such assessed systems (C1), as well as with initial research on using AI in safety-critical systems without formal coverage on assuring that such safety-critical AI is indeed safe (C2).

The second wave of research in which AI and safety have been jointly addressed is significantly more vigorous than the first one and comprises a trend of steep increase in all categories from 2016 onwards, except for outliers in 2018 and 2021 on C1. This indicates that further efforts on other areas – remarkably assuring that safety-critical AI reaches its safety requirements (C3) – have been increasingly investigated along with those already covered in the first wave.

In addition to the identified growth waves, another result worth identifying and analyzing is the drop in publications of all categories but C1 in 2021. Once such a decrease has occurred for a single year so far, it is deemed that it is still insufficient to characterize a consistent loss of interest in this category. It is also worth noting that C1 has still had significantly more publications in 2021 than in 2019 according to Figure 3, which suggests that, along with the increase of the other categories, further studies on AI and safety assurance are still relevant despite the aforementioned decrease of C1 in 2021. This is particularly true for research directly related to the safety assurance of AI-based systems – which bases all C3-categorized publications. C3 has actually reached a higher share among all categories in 2021 than in 2020 given its steeper increase in 2021 than the other categories with most published research (i.e., C1 and C2).

Finally, with regard to 2022 data, even though direct analyses are not feasible because of the restricted timespan of the preliminary results (up to August 26th), two trends can be identified. From a qualitative standpoint, it is noticeable that the effervescence on research joining AI and safety still persists, as 2022 data up to August 26th are comparable to the whole set of publications of 2019. Moreover, if one assumes the hypothesis that 2022 will follow the same publication rate observed up to August 26th (i.e., after 238 days since the year started), an estimate of the number of publications for 2022 can be obtained by multiplying the current 2022 results, depicted in Figure 3, by 365 days/238 days =1.53. By doing so and comparing the 2022 estimates with 2021, one can infer near-stability for categories C2 (97 vs. 99 respectively) and C5 (6 vs. 6, respectively), a 10% decrease on C3 (84 vs. 72, respectively), and steeper reductions on C1 (59 vs. 32) and C4 (11 vs. 2, respectively). This might indicate a lowered interest on C1 and C4, followed by a trend of continued interest on C2, C3, and C5. Since the latter three categories represent relevant themes towards full AI autonomy within safety-critical contexts, whereas the former two categories are closer to general-purpose applications of AI, such a behavior would not be unexpected if effectively confirmed.

B. C3 PUBLICATIONS Q-INDEX ANALYSIS

It is possible to expand on the bibliometrics of the 329 C3-classified publications by cross-analyzing them with the Q-index attributed to each of the C3 references. In order to improve the readability of the graphs used for this purpose, the fuzzy Q-index classification defined in subsection III-F (i.e., low, average and high) is herein adopted.

Among all the 329 C3 references, 115 (35.0%) were classified as low quality, 70 (21.3%) were classified as average quality, and 144 (43.8%) were classified as high quality. The rather significant quantity of low quality papers for C3 stems from the conservative C3 classification criterion explained in subsection III-D. By this criterion, some references were initially classified in C3 together with other categories because their TAK information was not significant enough to rule this classification out. After the full-text review, 73 of the 161 references jointly classified in C3 and in at least another category were deemed of low quality for C3. These 73 references represent 63.5% of the C3 low quality group.

Figure 4 shows how the yearly average Q-index has evolved with time from 1994 up to August 26th, 2022. This period has been defined because 1994 is the year in which the first C3 research paper has been published. Moreover, the period from 1995 to 2002 has been omitted from the graph to improve its readability because no C3 publications have been identified for any of these years. Moreover, the quantity of yearly C3 publications is explicitly listed on Table 1 to improve the understanding of the analyses.

It is possible to notice that, after two isolated peaks between 2003 and 2007 and on 2012, the Q-index has consistently increased with time during its 2016-2022 growing wave. The yearly average Q-index started with 1.0 on 2015 and has continuously grown up to 3.43 in 2022 except for two drops: one in 2017, when the growth wave was still at

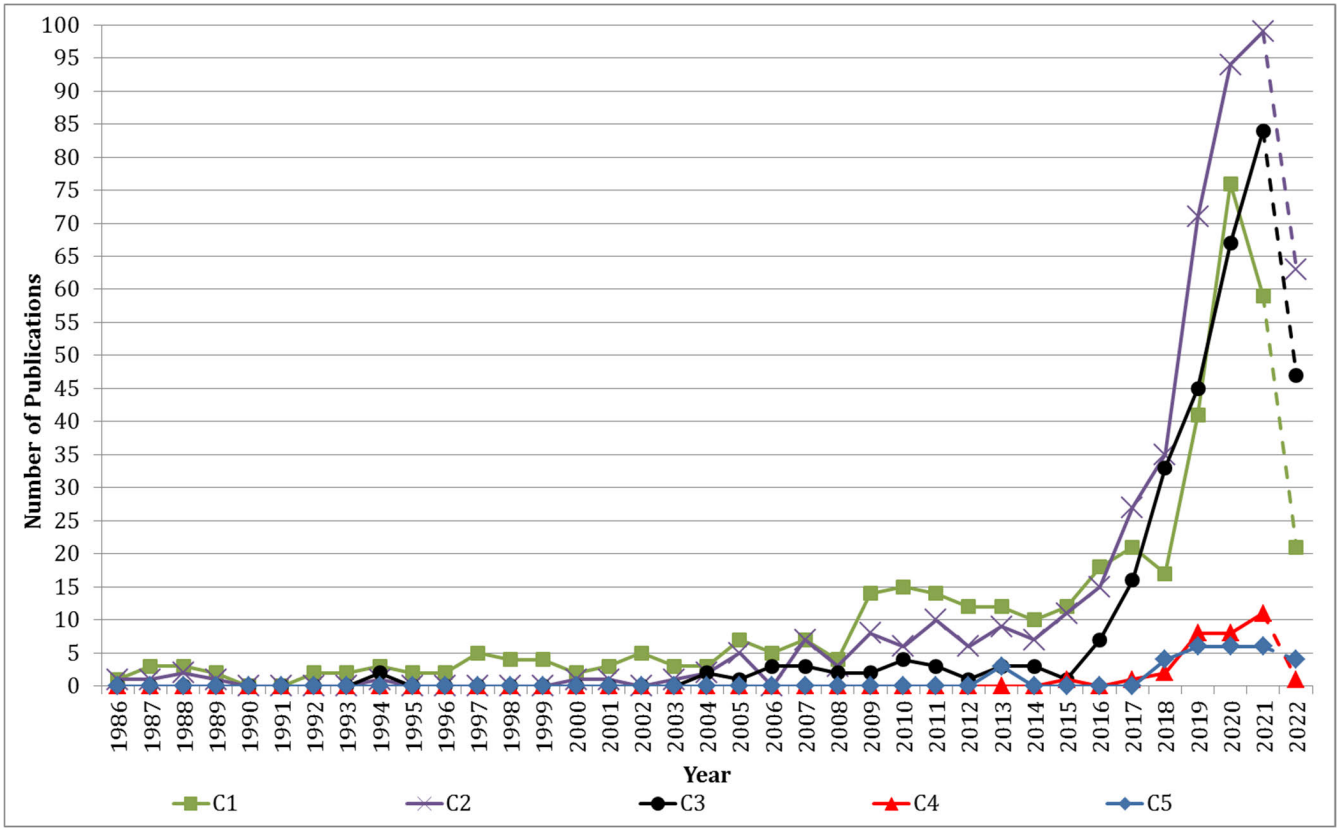

FIGURE 3. Evolution of the number of C1 to C5 references with time.

its beginning, and another one in 2021, with a small relative reduction of 7.6% in relation to 2020 (from 3.31 to 3.08).

It is worth highlighting that the isolated peaks between 2003 and 2007, as well as that on 2012, result from the fact that most of the few C3 publications on each of these years (no higher than 4 C3 publications, as per Figure 3) were deemed of high importance to the safety assurance of AI-based systems. This stems from early and successful attempts on addressing means to either assess if neural network-based control systems are safe [32], [33], [34], [35], [36], [37] or conceive fail-safe AI-based systems [38].

The remainder of the Figure 4 behavior can be understood by the analysis of Figure 5, which shows the relative growth of C3 references during the same period of Figure 4. Figure 5 shows, for each of the assessed years, the percentage of texts from each year which were rated with either a low quality Q-index (i.e., 0 or 1), an average quality Q-index (i.e., 2 or 3), or a high quality Q-index (i.e., 4, 5 or 6).

Up to 2015, when no more than 4 C3 publications were available per year, the percentage attributed to each of the aforementioned categories varies significantly. Starting in 2016, such oscillations have reduced due to the increase of C3 publications, and high quality publications have consistently risen in participation since then, with small drops on 2017 and 2021. Moreover, high quality C3 publications have become the most prevalent among all C3 references in 2019 up to 2022 (August 26th), yearly peaking at between 40% and 55% of the total of C3 publications at this timespan.

TABLE 1. Number of C3 publications per Year.

Year	Number of Publications	Year	Number of Publications
1994	2	2013	3
2004	2	2014	3
2005	1	2015	1
2006	3	2016	7
2007	3	2017	16
2008	2	2018	33
2009	2	2019	45
2010	4	2020	67
2011	3	2021	84
2012	1	2022 (up to Aug. 26th)	47

C. CONCLUDING REMARKS ON THE BIBLIOMETRICS ANALYSIS

The results presented in sections IV-A and IV-B are highly suggestive that the safety assurance of AI-based systems has

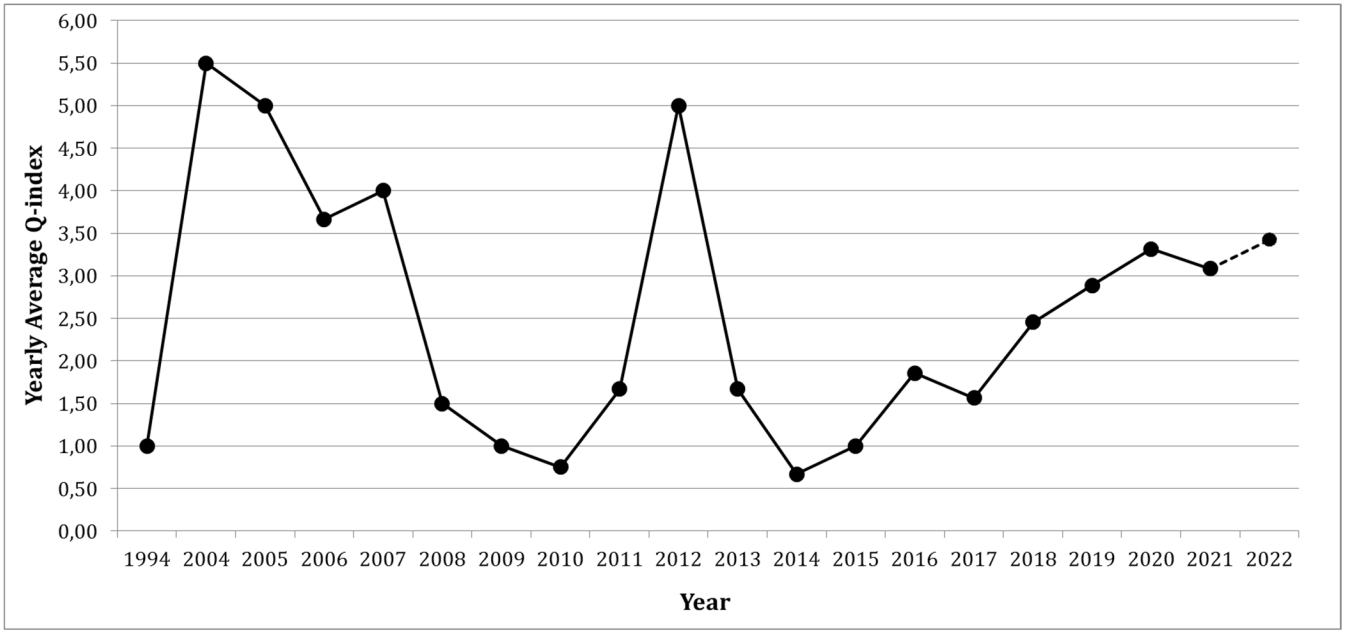

FIGURE 4. Evolution of C3 publications yearly average q-index with time.

been increasingly deemed worthy of relevant research by the research community especially from 2016 onwards. This reinforces the importance of the present work not only in compiling, assessing and reporting the progressively effervescent state of the art and future work raised by the research community on the safety assurance of AI-based systems, but also in expanding on these subjects by cross-fertilizing current research in order to draw further conclusions on these matters. Technical aspects regarding this 'expanded overview' on the state of the art and future work on the safety assurance of AI-based systems will be covered in the forthcoming sections of this paper.

V. STATE OF THE ART RELATED TO THE SAFETY ASSURANCE OF AI-BASED SYSTEMS

The objective of this paper section is to present the state of the art related to the safety assurance of AI-based safety-critical systems. It starts on subsection V-A with a brief introduction on how the relationship between AI and safety has evolved so far and up to the point when the safety assurance of AI-based systems became a major research problem on its own. Afterwards, the state of the art related to the safety assurance of AI-based systems per se is explored in subsection V-B.

A. RELATIONSHIP BETWEEN AI AND SAFETY: ORIGINS AND EVOLUTION UNTIL THE SAFETY ASSURANCE OF AI-BASED SYSTEMS

A summary of the overall evolution on how AI and safety have been combined with time is depicted in Figure 6, in which three major 'waves' of research are presented. These are detailed throughout this section.

The earliest records of research addressing the usage of AI in safety-critical applications date back to the mid-to-late-1980s and are related to using knowledge-based systems as a means to detect potential faults on nuclear power plants and report such faults to human operators. In this context, the information produced by the knowledge-based systems would support the decision-making process of human operators on triggering, e.g., preventive maintenance and emergency actions to respectively avoid and contain potentially unsafe scenarios [39], [40], [41], [42], [43]. Such trend of using knowledge-based systems to support human decision-making in safety-critical applications, which represents the 'first wave' shown in Figure 6, was still highly prevalent through the 1990s [44], [45], [46], [47], [48], during which only scarce efforts on other AI approaches, such as machine learning, have been carried out [49], [50].

On the early 2000s, a 'second wave', slightly stronger than the first, emerged with two major changes on the relationship between AI and safety. Firstly, machine learning techniques started replacing knowledge-based systems as the preferred AI technique used in research related to safety-critical systems. Secondly, efforts in including AI within the control loop of safety-critical systems, rather than just supporting the decision-making of human operators, also became increasingly more frequent.

One of the earliest research towards these changes is the one carried out by Wei [51], who presented an Artificial Neural Network (ANN)-based system that supports drivers of ground vehicles in performing safe lane-changing operations by supervised learning of potentially safe scenarios from video recordings of human drivers. Even though the system was still not developed aiming fully autonomous vehicles – which ultimately still makes it a human decision-making support system –, the results obtained by the author showed that his system was successful in mimicking human behavior in

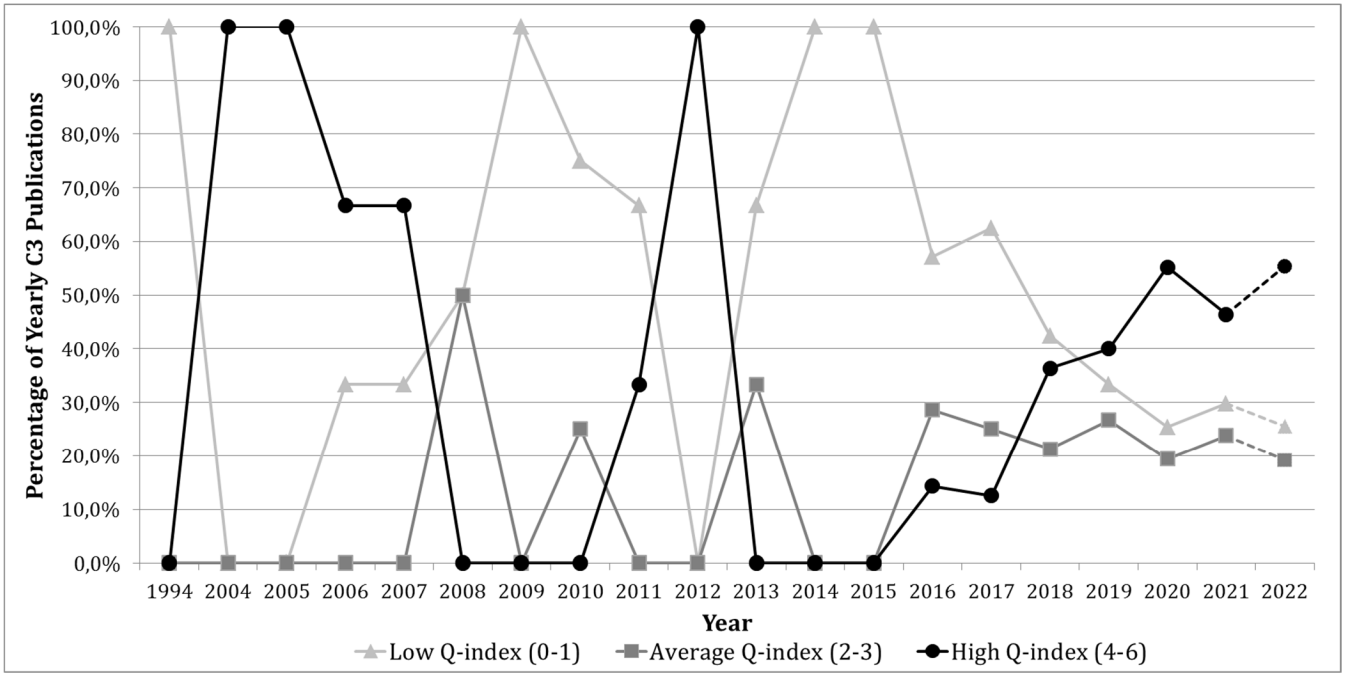

FIGURE 5. Relative evolution of yearly c3 papers per q-index fuzzy group with time.

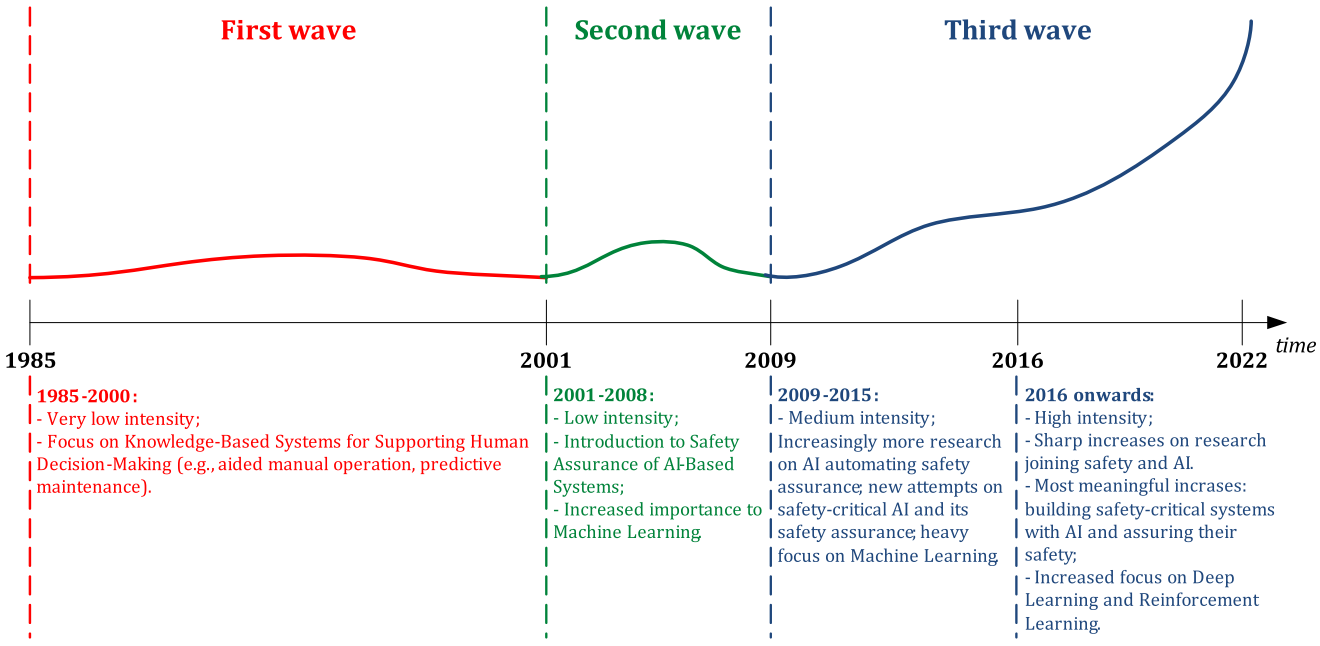

FIGURE 6. Overview of the research waves combining safety and AI.

safely recommending lane-changing operations yet retaining high driving performance (e.g., higher speed than state-of-the-art solutions by that time) while moving from one lane to another [51].

Another early study worthy of mention is the one by Kurd and Kelly [32], who have developed a white-box, fuzzy map-based model to design and represent ANNs used in safety-critical applications and which would be further successfully exercised within a Gas Turbine Aero-Engine system in the following years [33], [36]. With this effort, the authors have not only introduced the possibility of using AI as part of the control loop of safety-critical applications, but also discussed and exercised explainable AI, which would only emerge on its own as a concept and research theme within the AI field in the mid-2010s, based on its unconscious awareness at the expert systems era [52]. Moreover, the works by

Kurd and Kelly [32], [33], [36] represent the earliest studies that are highly relevant to the safety assurance of AI-based systems as per details covered in subsection V-B.

Since then, research that combines the areas of AI and safety assurance experienced significant effervescence. This trend, which characterizes the 'third wave' of research in Figure 6, emerged in 2009 as greater than the previous waves and became even stronger especially from 2016 onwards, as sustained by the bibliometrics analyzed in section IV. The main justification for the steep increase of publications involving AI and safety stems from the increase in cost-effective sensing and data processing capabilities of computer-based systems throughout the 2000s and 2010s, supported by parallel computing and, more recently, cloud computing [53]. With these features, computationally costly AI solutions – notably those related to deep learning –, have become feasible to pave the way towards implementing complex functions with AI [53] – including safety-critical ones.

So far, research involving AI and safety without directly addressing the safety assurance of AI-based systems (i.e., research included within categories C1, C2, C4, and C5, as per subsection III-D) have spanned a multitude of application domains and AI techniques. The target applications include, but are not limited to, power plants, transportation systems of several means, medical systems, process industries, and automating safety analyses. AI techniques, in turn, comprise a non-exhaustive list with several variants of ANNs (including deep learning – DL), logistic regression, k-nearest neighbors (kNN), decision trees (DTs), random forests (RFs), support vector machines (SVMs), boosting, and reinforcement learning (RL).

B. SAFETY ASSURANCE OF AI–BASED SYSTEMS: THE ROAD SO FAR

As introduced in subsection V-A, the series of research papers by Kurd and Kelly [32], [33], [36] has been considered the first meaningful efforts in exploring how to ensure that AI-implemented safety-critical functions indeed meet their related safety requirements. These were followed by 326 other research papers aiming to explore the safety assurance of AI-based safety-critical systems up to August 26th, 2022, with 144 of them[1] meeting the criteria for high relevance to the area, as explained in subsection IV-B.

The objective of this section is to expand on the safety assurance of AI-based systems' state of the art and contextualize it with the guidance of questions Q1 to Q4 from the SLR method (defined in subsection III-E). Special focus is given to the responses for these questions stemming from the 144 C3 papers that were deemed to highly contribute to the safety assurance of AI-based systems theme.

[1]This quantity includes the work by Kurd and Kelly [32], [33], [36].

1) OBJECTIVES OF RESEARCHING THE SAFETY ASSURANCE OF AI-BASED SYSTEMS – QUESTION Q1

By means of the SLR question Q1, the analysis of the C3 publications allowed identifying four mutually exclusive objective groups (OGs) related to their goals towards the safety assurance of AI-based systems:

- **OG1**: The research only aims to review and/or spot gaps on AI-based systems verification, validation and safety activities;
- **OG2**: The research aims to propose means to ensure that an AI-based system/function is safe and present supporting results;
- **OG3**: The research aims to apply methods defined in other research to ensure that an AI-based system is safe;
- **OG4**: The research covers other topics which may be either marginally related or unrelated to OG1, OG2 and OG3.

Table 2 summarizes the absolute and relative results of papers belonging to each of these OGs considering all C3 texts and only those rated with high quality as per their Q-index. The results indicate that high quality C3 references focus especially on OG2, which is expected given the scope of this research.

2) AI TECHNIQUES CONSIDERED IN THE SAFETY ASSURANCE OF AI-BASED SYSTEMS – QUESTION Q2

In order to evaluate what AI techniques have been indeed covered at the references on the safety assurance of AI-based systems, this information has been collected from each of the reviewed research papers by means of the SLR question Q2. During the analysis of the C3 publications, however, it has been noticed that some of them do not explicitly mention AI, and even those which do perform it with varying depth degrees with regard to AI variants, machine learning (ML) categories, and even specific AI and ML techniques. The observed variations are listed as follows:

a) AI Variants:
   i. No explicit mention to AI;
   ii. AI in general;
   iii. Machine Learning in general (ML);
   iv. 'Classic AI' search, game theory and evolutionary algorithms;
   v. Knowledge-Based Probabilistic Models (KBPMs), such as Bayesian approaches, Kalman and Particle Filters, and Dempster-Shafer Theory.

b) ML categories:
   i. Supervised Learning (SL);
   ii. Unsupervised Learning (UL);
   iii. Reinforcement Learning (RL);
   iv. Deep Learning (DL).

c) Specific AI and ML techniques:
   i. Artificial Neural Networks (ANNs);
   ii. Decision Trees (DTs) and Random Forests (RFs);
   iii. Support Vector Machines (SVMs).

TABLE 2. Summary of question Q1 results.

		OG1	OG2	OG3	OG4
All C3	Absolute	27	179	29	94
	Relative	8.2%	54.4%	8.8%	28.6%
High Quality C3	Absolute	17	123	4	0
	Relative	11.8%	85.4%	2.8%	0.0%

The obtained results are summarized in Table 3 for both the entire set of C3 references and the subset of those whose Q-index is within the high quality fuzzy group.

Firstly, it is worth noting that the aforementioned groups of AI variants, ML categories and specific AI and ML techniques are non-mutually exclusive, meaning that each C3 reference can be within more than one of these qualifying groups. Hence, summing the references of all groups in Table 3 yields results that are greater than the actual quantity of assessed references (i.e., greater than 329 for the entire set of C3 references and greater than 144 for the set of high quality C3 references).

Moreover, it is possible to notice that the proportion of high quality references which do not explicitly mention AI is steeply smaller than that of the entire set of C3 publications (20.1% vs. 3.5%). This result corroborates that relevant efforts on the safety assurance of AI-based systems shall explicitly address AI somehow, which occurs with 139 out of the 144 (96.5%) high quality C3 references.

Furthermore, the higher-graded references also focus especially in ML (>35%), remarkably based on ANNs (>50%). In addition to this, these which explicitly describe the ML approach allow checking research trends towards DL (>22%), RL (>18%), KBPM (9%) and SL (>8%), with DL and SL also intimately related to ANNs. When adding to this analysis the graph of Figure 7, which shows the yearly proportion of C3 references for each AI variant, ML category and AI and ML technique from 2018 onwards, there is a rising trend of publications specifically on ANNs and a decreasing trend of ML in general, with ANNs overtaking ML in general from 2020 onwards as the main theme of research papers. DL has oscillated on the same period of time between 16% and 33% of yearly representativeness, tying with ML in general on 2021, but on 2022 (up to August 26th), it has lagged behind it. RL has also oscillated during the same timespan; nevertheless, despite reaching its lowest representativeness on 2019 (<6%), it has significantly grown in relevance on the forthcoming years, peaking at more than 38% on 2022 (up to August 26th).

The overall prevalence of ANNs, DL, and RL suggests two main characteristics. On the one hand, exploring the safety of ANNs, DL, and RL follows the current AI area trend in using them because of their flexibility in providing somewhat accurate models for functions which are hard to be precisely specified [54], [55] and that can ultimately benefit from exploring and exploiting an operational environment for fine-tuning prior to revenue service [56]. On the other hand, a significant amount of safety assurance research is directed towards analyzing rather opaque and inherently

TABLE 3. Summary of question Q2 results.

AI Reference	All C3		High Quality C3	
	Absolute	Relative	Absolute	Relative
Non-Explicit IA	66	20.1%	5	3.5%
AI in General	63	19.1%	30	20.8%
ML in General	78	23.7%	51	35.4%
Classic AI	14	4.3%	3	2.1%
DL	55	16.7%	32	22.2%
RL	48	14.6%	27	18.8%
SL	18	5.5%	12	8.3%
UL	13	4.0%	2	1.4%
ANN	127	38.6%	73	50.7%
SVM	10	3.0%	5	3.5%
DT/RF	13	4.0%	7	4.9%
KBPM	27	8.2%	13	9.0%

hard-to-explain ML variants [57], [58], [59], whereas answers to deal with the safety assurance of simpler and explainable AI models, such as DTs, are nonetheless scarce.

For instance, the research of Hernández-Orallo et al. [60] and of Groza et al [61] were the single high-quality efforts explicitly related to developing safety-critical AI with DTs. Even if all publications within at least one of the groups 'AI in General' and 'ML in General' were sufficiently generic to be applied to simpler and explainable AI models, these would still account for at most 45.1% of all publications (65 out of 144 publications). Such result is lower than the proportion of all research papers specifically devoted to the safety assurance of either ANNs or DL (54.2%; 78 out of 144 publications), and even lower to that when RL is included along with ANNs or DL (61.8%; 89 out of 144 publications).

Finally, most of the 2022 representativeness increase of older, more 'traditional' AI models, such as DT/RF and KBPM, is due to the increasing number of relevant literature reviews covering them [19], [24], [25]. Such reviews have been cited and compared to this SLR in section II. The exceptions are the research efforts by Ruchkin et al. [62], with reference to Bayesian models and logistic regression, Groza et al. [61], with random forests and rule-based systems, Musau et al. [63], with rule-based systems, and Bai et al. [64], with random forests.

3) METHODS FOR THE SAFETY ASSURANCE OF AI-BASED SYSTEMS AND THEIR RESULTS – QUESTIONS Q3 AND Q4

As per Figure 8, the full-text review of C3 references allowed identifying five significant approaches in order to deal with the safety assurance of AI-based systems. These approaches are based on (i.) performing extensive black-box testing

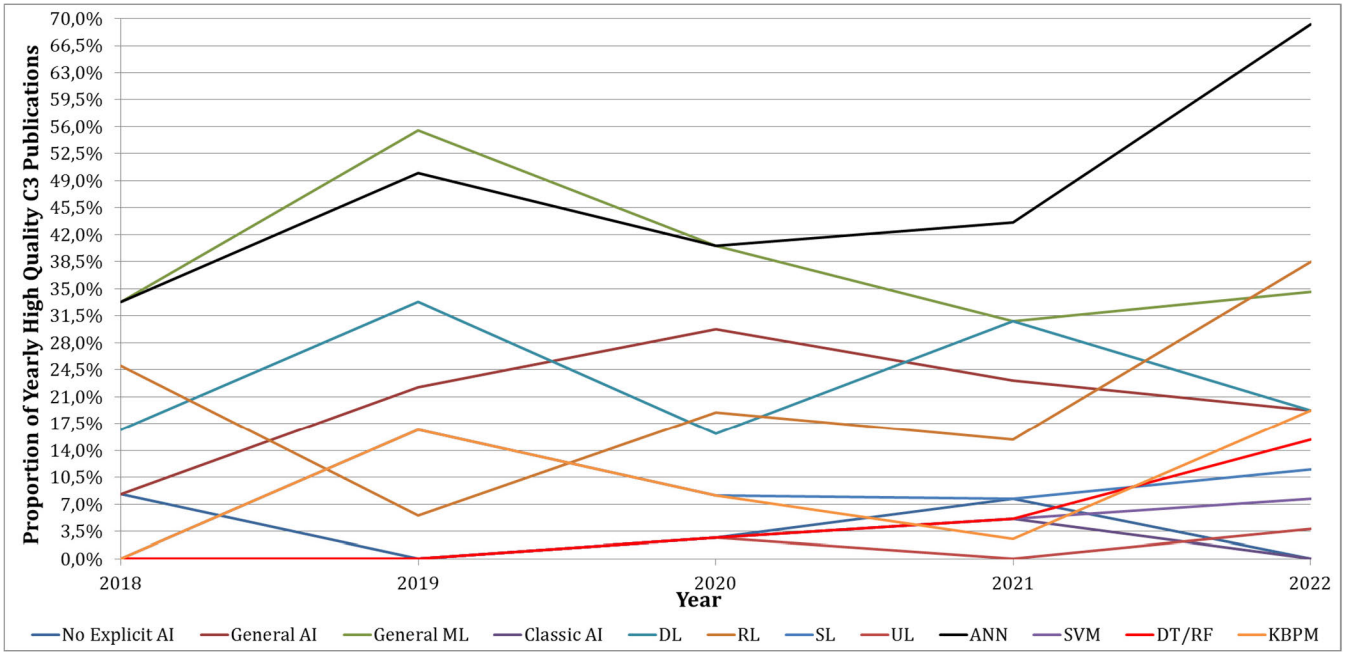

FIGURE 7. Proportion of yearly C3 high quality publications between 2018 and 2021 for each AI Variant, ML Category and AI and ML technique.

of AI, (ii.) utilizing safety envelopes to limit the response of the AI to a safe image set, (iii.) crafting fail-safe AI, (iv.) combining explainable AI with white-box analyses to provide in-depth understanding of the underlying AI models, and (v.) conceiving systems-level methods and processes, potentially merging the four other approaches, to systematize the safety assurance of AI.

The characteristics of each of these approaches are presented, along with the main results of their most relevant research papers as per the Q-index attributed to them, in the following subsections. All information herein presented has been captured by means of questions Q3 and Q4 of the SLR questionnaire.

a: SAFETY ASSURANCE BASED ON BLACK-BOX TESTING

The first approach is related to specifying and performing **test cases** as exhaustively as possible (in simulated or real world scenarios) so as to try to address the most variations of AI-based systems behaviors whilst **treating AI as a black box** due to its complexity. This would be achieved by different input stimuli (i.e., validation datasets), which would then translate into exercising different paths of the AI model by such inputs. Hence, the main advantage of this approach is that dealing with AI as a black box allows abstracting the underlying difficulty in understanding its internal characteristics, thus making it simpler and faster to reach effective results to either support or deny that a system is safe [65].

A summary of the advantages and disadvantages of the black-box testing approach, further discussed throughout this subsection of the paper, is presented in Figure 9. Green text boxes indicate potential advantages of the approach, whereas red text boxes indicate its disadvantages and difficulties.

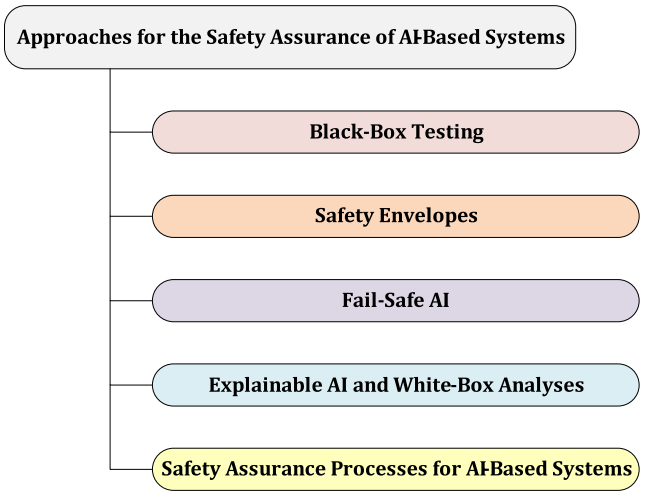

FIGURE 8. Approaches for the safety assurance of AI-Based systems.

This approach has been explored, e.g., on the research papers by Meltz and Guterman [66], [67], Watanabe and Wolf [68], Sun et al. [69], and Hussain et al. [70], all of which within the context of using AI on safety-critical functions of Unmanned Ground Vehicles (UGVs). In addition to them, the literature reviews by Tahir and Alexander [21] and by Corso et al. [65] are also devoted to the safety assurance of AI-based systems by means of black-box testing. Finally, the research published by Kozal and Ksieniewicz [71] and by represents an example of the black-box testing approach applied to safety-critical medical systems.

Meltz and Guterman [66], [67] and Watanabe and Wolf [68] have developed UGV models in which AI is responsible for performing safety-critical functions, such as braking and

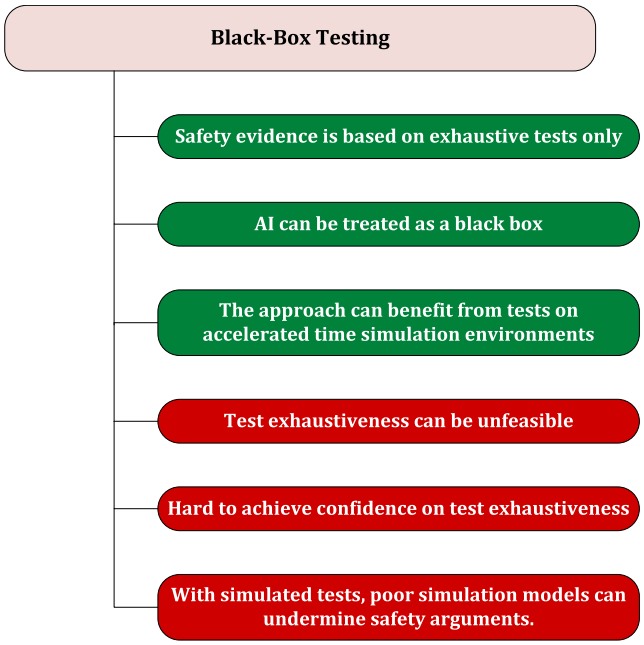

FIGURE 9. Advantages and disadvantages of black-box testing.

collision avoidance. The strategy presented in all studies for validating their models is based on black-box testing of the AI modules, combining simulations and field tests. The main conclusions of such studies are as follows:

a) Meltz and Guterman [66], [67] have performed simulations and physical tests of an UGV travelling on a 100m-long pathway with obstacles and have detected that their black box testing strategy was able to detect nearly 1% of unsafe scenarios, in which the UGV collides with obstacles. This rate is deemed overly high for certifying a safety-critical system with typically restrictive safety requirements;
b) Since Watanabe and Wolf [68] have only presented an UGV design and testing framework, no practical results from their conceptual methods have been covered;
c) Both groups of researchers have anticipated that their black-box testing strategies are insufficient for assuring that the AI-dependent functions are sufficiently safe, and that they shall only be considered as a starting point towards assuring safety [66], [67], [68].

Sun et al. [69] proposed using supervised learning AI to reduce the efforts in identifying relevant test cases and, hence, maximize the coverage of black-box tests applied to safety-critical AI elements within a specific timespan. Even though Sun et al. [69] were able to obtain reasonably positive results towards their objective on case studies involving UGVs, with 99% effort reduction whilst retaining a 90% coverage of safety-critical tests with a confidence rate of 90%, two drawbacks have been identified.

Firstly, the authors themselves claim that the models employed on the case study are not sufficiently faithful for real applications and that further study is still needed on the validity of the results they obtained [69]. Secondly, even if such results are confirmed, both coverage and confidence rate indexes that were obtained are deemed insufficient for certifying a safety-critical system with typically stringent safety requirements.

Still on the UGV area, Hussain et al. [70] have developed and presented a safeguard for autonomous driving systems called DeepGuard, whose objectives are (i.) to check whether the driving context might violate safety requirements and (ii.) enforce the needed safe actions if such a situation is detected. DeepGuard is responsible for modeling the driving context as a first-order time series and inputting it to an autoencoder (i.e, an ANN whose aim is to reproduce its inputs on its outputs), which is able to detect whether there are major differences to the operational environment that can lead the UGV to deviate from its expected safe behavior. The contribution of DeepGuard to overall safety is assessed for two UGV control functions – namely, collision avoidance and lane changing – following a black-box testing approach based exclusively on metrics extracted from a confusion matrix (e.g., precision, recall, f1-score) [70].

Tahir and Alexander [21] have presented a literature review on means to certify safety-critical AI-based UGVs by means of black-box testing results. On the one hand, relevant techniques to improve the coverage of black-box tests have been identified, such as High Throughput Testing (HTT), Search-Based Software Testing (SBST), and pseudorandom test case generation, One the other hand, Tahir and Alexander [21] have stated that most of their reference studies were deemed of low quality, and that the lack of extensive test coverage maximization techniques is still an obstacle towards using black-box testing as the sole tool for assuring that a safety-critical system is indeed safe.

Corso et al. [65] have also explored improving the coverage of black box techniques to support the safety validation of AI-based systems, but it differs from the review by Tahir and Alexander [21] because Corso et al. [65] focused on using AI itself to increase the coverage of black-box testing. The main strategies discussed by the authors is that AI variants for optimization, planning, and reinforcement learning are relevant tools in covering a wider range of black-box tests whilst reducing the efforts to reach this coverage. The authors, however, do not discuss how to ensure that such AI tools themselves are sufficiently safe to guide the black-box tests of safety-critical AI-based systems.

Finally, Kozal and Ksieniewicz [71] have indirectly utilized black-box testing of AI as the prime approach to assess whether the AI used in their study behaves in a safe way. The objective of their research is to develop a system that is able to classify whether heartbeats are healthy or of four different types of arrhythmias by using Residual Neural Networks (ResNets) to analyze and classify 187-sample time series that represent the input heartbeats. The authors have developed means to reduce the heavy imbalance of the input datasets towards healthy heartbeats and compared and contrasted the approaches by using black-box tests, performance metrics such as category-specific precision and recall, and statistical

tests. Even though the results indicate that the imbalance reduction techniques have improved safety with high certainty levels [71], the authors have not focused on rigorously defining safety requirements and assessing whether the results are sufficiently appropriate to support the diagnosis of heart diseases.

In addition to the criticisms and limitations identified on the previous reviewed references, several other researchers (e.g., Harper and Caleb-Solly [72], Koopman and Wagner [73], Musau et al. [63], and Wu et al. [74]) claim that no sufficiently exhaustive tests can be carried out for safety-critical systems in due time given the strict requirements to which they must comply (e.g., failure rates lower than 10^{-8} failures/hour for highly safety-critical systems in continuous operation, as per requirements derived from IEC61508 [9]). In order to circumvent this, four other approaches for the safety assurance of AI-based systems have been recently explored. They are covered in the next subsections.

b: SAFETY ASSURANCE BASED ON SAFETY ENVELOPES

The basic idea of this approach is to restrain the behavior of AI-based systems by design within a **deterministic** (i.e., non-AI-implemented)**safety envelope** (alternatively referred to as **safety cage**). In this context, such an envelope constrains the overall system response to a knowingly safe image set by design regardless of its AI, thus leading the underlying AI elements to play at most a minor role on safety. As a result, typical safety assurance methods for non-AI-based systems would suffice, since these would be solely applied to the non-AI-related safety envelope elements [18].

An overview of the advantages and disadvantages of the safety envelope approach, thoroughly reviewed in the remainder of this section along with representative references, is presented in Figure 10. Green text boxes indicate potential advantages of the approach, whereas red text boxes indicate its disadvantages and difficulties.

Solutions of this category have been presented and discussed by, e.g., Machin et al. [75], Shafaei et al. [76], Kuutti et al. [77], and Lazarus et al. [58]. Further information on each of these research papers is presented henceforth.

Machin et al. [75] have presented a general framework to translate safety requirements into predicate logic rules that formally define safety envelopes for active safety monitors. A case study of a mobile manipulator robot for co-working led to only partially successful results in defining safe constraints for the robot operation, with the following limitations:

a) Some safety requirements could not be addressed by means of the proposed framework due to the lack of observable data to generate safety envelopes [75];
b) Physical tests evidenced that the generated safety envelopes still allowed violating some safety requirements [75];
c) There is no explicit mention as to whether AI has indeed been used within the case study robot design.

Shafaei et al. [76] have crafted a set of recommended actions to reduce the impacts of the underlying uncertainties of machine learning for safety-critical components used in UGVs. Among the recommended actions, the authors highlight creating ontologies to enforce design level decisions and translate them into a safe envelope that limits the response of ML-based components. Since the research paper solely focuses on presenting the proposed method for dealing with ML-related uncertainties, no practical results have been obtained by the authors [76].

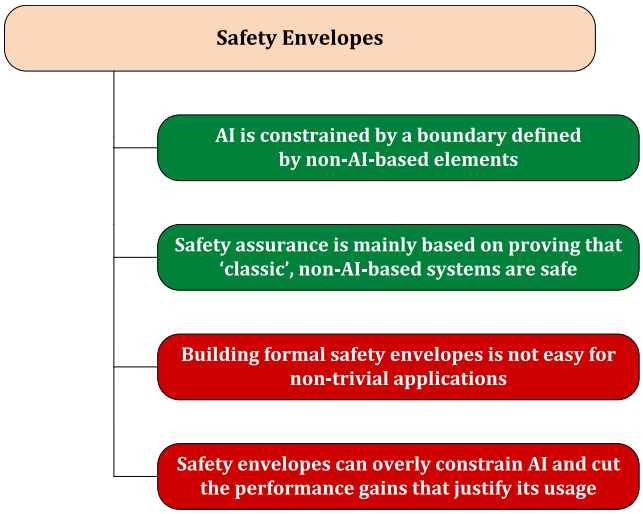

FIGURE 10. Advantages and disadvantages of safety envelopes.

Kuutti et al. [77] have explored implementing redundant safety envelopes on the control loop of an UGV so that the safety envelopes avoid front collisions based on two AI-based movement controllers: a Deep Neural Network (DNN)-based controller for optimum performance, and a suboptimal ANN-based controller with less layers. Within a simulated environment, Kuutti et al. [77] have observed that the safety envelopes prevented unsafe scenarios and that such safe action was required with a higher frequency when the suboptimal ANN-based controller was in charge of controlling the UGV instead of the DNN-based one.

Lazarus et al. [58] have developed an RL approach to synthesize safety envelopes which aims to increase their flexibility by means of dynamic boundaries determined according to operational characteristics. The authors have presented positive results of their result in simulated case studies involving Unmanned Aerial Vehicles (UAVs), since none of the simulated UAVs went outside their respective safety envelopes even when adverse operation conditions (e.g., strong winds) were exercised. Nevertheless, since RL determines the synthesized safety envelopes, assessing *a priori* if underlying RL models are sufficiently safe is still needed to ensure that the safety envelopes synthesized with it are themselves indeed safe. This last aspect has not been discussed by Lazarus et al. [58].

Safety envelope-based solutions are mostly criticized for two main reasons: (i.) underlying difficulties in formally defining them, and (ii.) their inherent feature of overly constraining the performance gains that AI can introduce [78].

The approaches presented in the next subsections aim to somehow address these limitations.

c: SAFETY ASSURANCE BASED ON FAIL-SAFE AI

Another approach towards the safety assurance of AI-based systems comprises research whose aim is to improve the **architecture** and the **learning process of AI-based systems** so that they approach **fail-safe characteristics**.

A graphical abstract of the advantages and disadvantages in using fail-safe AI, as discussed more comprehensively throughout this section of the paper, is presented in Figure 11. Green text boxes indicate potential advantages of the approach, whereas red text boxes indicate its disadvantages and difficulties.

The main advantage of the fail-safe AI approaches is the soundness of the resulting safety assurance arguments, since they are typically supported by semi-formal or formal analyses with strong mathematical background.

Gillula and Tomlin [38] have developed the GSOLR (Guaranteed Safe Online Learning via Reachability) scheme, whose aim is to define rules that restrict the AI behavior on potentially unsafe boundaries of the AI image set. The pre-condition for extracting these rules is determining the AI image set, which is carried out by means of Hamilton-Jacobi-Isaacs reachability analyses.

A real case study on a quadcopter that shall track a ground vehicle using a fail-safe AI tracking system has been presented by the authors. Even though its results have supported the soundness of the authors' method, since the quadcopter did not lose track of the ground vehicle, the quadcopter was not subject to challenging situations such as the need to deviate from in-course obstacles [38].

Jaeger et al. [79] have developed a fail-safe, model-based reinforcement learning scheme whose aim is to create a dynamic safety envelope – therein referred to as a 'Region of Safety (RoS)' – to facilitate the design of safety-critical self-adaptive and cooperative multi-agent systems. The authors have exercised their method by means of simulated case studies in which UGVs cooperatively adapt their learning-based cruise control systems to avoid potentially unsafe situations as they travel towards their objectives. The final results support that the method developed by the authors is safe, since no potentially unsafe situations were identified throughout the case studies. However, the authors themselves claim that underlying data and model uncertainties were simplified and restrained to low levels of Gaussian noise, which could ultimately undermine the validity of their safe results in real-world scenarios [79].

Lin et al. [80] have presented the Abstraction Refinement-Guided Training (ART) as a formal methods-based means for building correct-by-construction ANNs by minimizing a loss function which quantifies the learning errors of the ANNs with time. The authors have exercised ART by means of two simulated case studies: one considering 45 different ANNs to control an UAV anti-collision system, and another one based on an anti-collision system for UGVs. The results

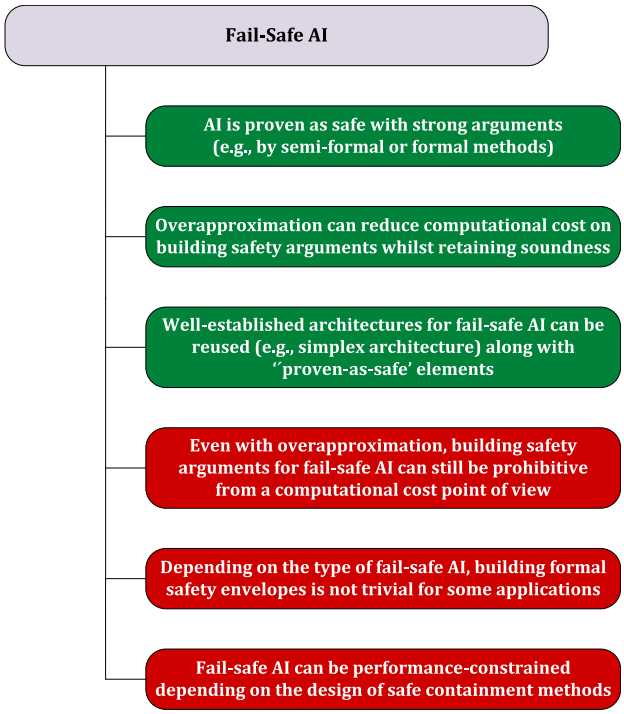

FIGURE 11. Advantages and disadvantages of Fail-Safe AI.

presented by the authors, however, have revealed limitations on the feasibility of fail-safe ANNs with ART on the UGV anti-collision system, since potentially unsafe situations were still identified after ART has been applied [80].

Zhao et al. [81], Zhao et al. [82], and Peruffo et al. [83] have developed means to synthesize intrinsically fail-safe ANNs by means of barrier certificates, so that formal boundaries for safe and unsafe state sets for the ANNs are explicitly established. Mixed results were obtained with this approach. One the one hand, Zhao et al. [82] presented a hypothetical case study in which not all potentially unsafe situations could have been avoided by applying their method. On the other hand, Zhao et al. [81] and Peruffo et al. [83] had positive safety results. The UGV control application covered by Zhao et al. [81] led to satisfactory safety-related results, whereas Peruffo et al. [83] resorted to case studies based on hypothetical benchmark systems of up to eight dimensions, which have also led to results that corroborate safety requirements have been satisfied.

Sha et al. [84], Claviere et al. [85], and Wang et al. [86] have presented approaches in which the safe response of ANNs is overapproximated, i.e., overestimated. The main objective of overapproximation is to reduce the computational cost of the underlying exact solutions on defining a safe set for an AI model whilst obtaining an overapproximated safe state set in which the actual exact safety task is contained. With overapproximation, an output is certainly safe when it is within the overapproximated safe set, whereas no conclusions can be drawn on its safety when outside it [84], [85], [86]. In the latter case, additional safety policies or analyses shall be applied [85].

Sha et al. [84] have explored their proposed overapproximation scheme by simulating a mass-spring damper system controlled by a DNN and varying the latter's architecture (e.g., number of neurons, number of hidden layers, and activation functions) in each experiment. With the aid of a prototyping tool of their own overapproximation model, the authors were able to create 10 safe DNNs out of the 12 exercised architectures [84]. Since at least a single solution is sufficient to deal with a specific problem, it is possible to consider that the authors were successful in crafting a proven-as-safe AI-based solution for their case study.

Claviere et al. [85], in turn, have developed means to overapproximate ANNs strictly based on the ReLU (Rectified Linear Unit) activation function. Their approach has been exercised with a case study that involves a modified version of the Airborne Collision Avoidance System for Unmanned Aircraft (ACAS Xu), in which ANNs were introduced as a means to generate safe UAS maneuvers whilst reducing the storage needs of the original ACAS system. The case study scenario, involving two aircraft in potentially conflicting routes to be resolved by the action of ACAS Xu, showed that the overapproximation of the ANNs yielded to safe situations in 98.8% of the tested settings, whereas the remaining 1.2% could not be proven as safe (i.e., they are not necessarily unsafe, but there is no sufficient evidence to say otherwise).

Wang et al. [86] have also focused their efforts on ReLU-based ANNs with three main objectives: (i.) tightening the overapproximations more than reference studies, (ii.) improving the overall processing time in overapproximating neural networks, and (iii.) incorporating underlying uncertainties of input data into the overapproximation calculations. By means of a case study of an advanced cruise control system for UGVs, the authors have reached the overapproximate set of its ANN and the conclusion that such an overapproximation would be safe if the uncertainties of inputs were constrained to a specific interval. When exercising the ANN with input data within and outside such an uncertainty range, the authors have obtained proof to support the soundness of the previous conclusion: all outputs of the ANN were safe when the uncertainty input bounds were respected, and unsafe states when reached when this condition was not met [86].

Other research in which tools for formally verifying ANNs are covered are the ones by Zhu et al. [87] (ReachNN – Reachability of Neural Networks), Ivanov et al. [88], [89] (Verisig and Verisig 2.0), Sidrane et al. [90] (OVERT[2]), Tran et al. [91] (NNV – Neural Network Verification), Fahmy et al. [92] (HUDD – Heatmap-Based Unsupervised Debugging of Neural Networks), Pulina and Tacchella [93] (Neural Networks Verifier – NeVer), and Katz et al. [94] (Marabou). The DeepCert tool by Paterson et al. [95] mixes formal verification of ANNs and DNNs used in image processing functions with black-box tests that aim to model potential image corruptions due to e.g., haze, blur and contrast changes. Moreover, other tools that can aid the verification of discrete-time systems with AI, such as dReal, dReach, and Flow $*$, have also been covered in the research by Tuncali et al. [96] and Val et al. [97].

Phan et al. [98] and Shukla et al. [99] have proposed architectural models which define the so-called simplex architecture for safety-critical ML-based systems. This architecture comprises four main modules: (i.) an AI controller and three non-AI-based elements: (ii.) a reference controller, which has been proven to safely accomplish the same safety-critical function of the AI controller albeit with subpar performance, (iii.) a safety-critical controller switch, which chooses the output of the AI controller if it is safe or the output of the reference controller otherwise, and (iv.) an optional AI controller adapter, which improves the AI controller with time by means of a learning process whenever it produces an incorrectly permissive (unsafe) output. The objective of the preexistent non-AI-based safe elements is twofold: (i.) ensuring safety when AI-based controllers fail to do so, and (ii.) leveraging the runtime learning of AI-based controllers so that the safe controller is used as little as possible with time.

Phan et al. [98] have exercised the full simplex architecture by means of two simulated case studies with ANN-based AI controllers: a moving-target tracking system for UGVs and an automated insulin pump for medical patients with diabetes. In both scenarios, the simplex architecture as a whole led both systems to behave safety: the UGV was able to track a moving target and avoid colliding with it, and the insulin pump avoided long-term hyperglycemia and short or long-term hypoglycemia [98].

Shukla et al. [99], in turn, designed an ANN-based control system for UAVs based on the simplex architecture, albeit disregarding the AI controller adapter on their model. Case studies carried out in simulated environment and in hardware-in-the-loop scheme (i.e., with the physical implementation of the UAV control system) supported that the UAV control system met its safety requirements, since no collisions with other elements have occurred. Furthermore, the authors have also observed proper switching between the AI controller and the reference controller whenever needed to avoid unsafe scenarios [99].

In addition to Phan et al. [98] and Shukla et al. [99], other recent research has explored the usage of the simplex architecture for safety-critical systems with AI. On 2022, for instance, four research papers report its usage: Chen et al [100], Peng et al. [101], and Wang et al. [56] have used the simplex architecture to support safety-critical functions on UGVs, whereas Thumm and Althoff [102] have experimented its usage on industrial environments with human-robot collaboration.

In all three UGV-related research papers, the authors have conceived a RL-based AI controller and a proven-as-safe reference controller to perform driving control functions. It is important to highlight that, whereas Phan et al. [98]

[2]A formal definition of the acronym 'OVERT' is missing in its originating reference by Sidrane et al. [90].

and Shukla et al. [99] have crafted a non-AI-based reference controller as the reference controller, Chen et al [100], Peng et al. [101], and Wang et al. [56] opted for using AI controllers which have been proven-as-safe by means of overapproximate mathematical models. In all three research papers, the authors have performed simulated case studies in which the safety-critical UGV functions they explore are on controllers included in the simulation loop and reached overall positive conclusions with regard to safety assurance whilst also retaining adequate performance.

Finally, Mehmood et al. [103] has extended the original simplex architecture by adding to it a look-ahead mechanism which loosens the safety requirements of the reference controller, allowing the latter to be also AI-based whilst ensuring global system safety. In the approach proposed by the authors, the safety-critical controller switch is augmented with two capabilities: firstly, it is able to process the immediate-future safety states of the whole system; secondly, it carries out reachability analyses on the reference controller to check whether it will reach or not the near-future safe states. If safety is not ensured, two safe actions are possible: (i.) the reference controller downgrades to previous versions up to meeting the safety constraints, or (ii.) the augmented safety-critical controller switch takes a deterministic safety decision if the downgrading of the reference controller times out.

The authors have exercised the extended simplex architecture with two simulated case studies: a model-predictive control for multi-robot coordination, and a collision avoidance mechanism for aircraft. The authors have not identified potentially unsafe scenarios from a systems point of view, but highlighted the difficulty in using their extended simplex architecture because it requires a significant amount of storage space for the look-up tables of the reference controller (e.g., hundreds of gigabytes for the aircraft collision avoidance controller) [103].

The main limitations on designing fail-safe AI is the high computational cost of reachability analyses even with simple, non-deep AI models with not many input variables. This is due to the inherently NP-Hard computational complexity of the involved models, which require techniques such as Satisfiability Modulo Theories (SMTs) and Linear Programming to be solved [104]. Furthermore, even if simplification schemes such as overapproximations are considered, these can either mask potential safety issues if misconceived, or even lead the resulting system to be 'excessively safe', to the point that an allegedly better performance introduced by the AI might be unjustified by the added complexity [74], [82], [87].

Moreover, when fail-safe architectures make use of non-AI-related fail-safe elements to mitigate potentially unsafe responses of the AI, they also share the same corresponding limitations of safety envelopes. On the other hand, though, as the AI elements are designed to learn with the previous unsafe responses, greater flexibility can still be achieved than with safety envelopes per se. Finally, since there is still no consensus on which types of AI redundancy support the design of fail-safe, no 'design patterns' towards fail-safe AI architectures have been established so far [37].

d: SAFETY ASSURANCE BASED EXPLAINABLE AI AND WHITE-BOX ANALYSES

Explainable AI (XAI) is also deemed an emergent topic to address the safety assurance of AI-based systems, since it aims to build AI elements which clearly allow humans to identify decisions taken by AI and their underlying reasoning. This ultimately makes it easier to assess AI-based systems with **white-box analyses**, which are the norm of traditional approaches used with non-AI-based safety-critical systems, and also allows building robust safety arguments due to the in-depth analyses [105].

A summary of the advantages and disadvantages of the approach combining XAI with white-box analyses, discussed in more detail throughout this subsection of the paper, is presented in Figure 12. Green text boxes indicate potential advantages of the approach, whereas red text boxes indicate its disadvantages and difficulties.

Kurd et al. [37] and Kurd and Kelly [32], [33], [36] have developed a W-shaped systems lifecycle to build and analyze safety-critical hybrid ANNs based on Fuzzy Self-Organizing Maps (FSOMs). The lifecycle introduces the concept that the aforementioned hybrid ANNs can be assessed as safe because they are explainable. Such explainability results from the ANN being generated by a gradual refinement of the data used in the ANN learning as its design progresses, which led these ANNs to be called Safety-Critical Artificial Neural Networks (SCANNs). This refinement, in turn, is automatically achieved by the FSOMs, which are created by human experts through fuzzy rules that explicitly define the ANNs expected behavior.

A case study of an hybrid, FSOM-based ANN to control a gas turbine has been presented by the authors along with results that support that the system is explainable and safe for the three safety requirements they defined – namely, (i.) avoiding engine surge, (ii.) avoiding turbine blade overheating, and (iii.) avoiding engine overspeed [37].

Grushin et al. [105] have conceived an overapproximation model to translate Long Short-Term Memory (LSTM) ANNs into explainable models by clearly defining hyperplanes which characterize the image set of the LSTM ANNs. The case study explored by the authors has aimed to conceive and assess the safety of an explainable LSTM ANN-derived model which is in charge of predicting if an aircraft will reach a degraded state. In this context, the LSTM ANN decides whether a degradation is expected based on both the aircraft's internal systems' health and the operational context of the global airspace as monitored by the aircraft itself [105].

The authors have presented two main results. Firstly, the hyperplanes which define the boundaries between operational and degraded states corroborate the explainability of the model. Secondly, the example of the case study leaned towards safety, as the explainable model derived from the

FIGURE 12. Advantages and disadvantages of explainable AI with white-box analyses.

LSTM ANN tended to predict degradations earlier than expected [105].

Salay et al. [106] have presented a Failure Modes, Effects and Criticality Analysis (FMECA) approach crafted to the safety analysis needs of AI-based systems. The base of their method is that four sources of AI failure modes shall be considered: (i.) failures on abstracting real world elements by AI, (ii.) AI uncertainties, (iii.) susceptibility to adversarial attacks, and (iv.) AI failures on dealing with the compromise between safety and other requirements (e.g., performance).

The proposed method has been applied to a ML-based UGV system which is in charge of either maintaining the vehicle within its traffic lane or moving to another lane should there be obstacles and safe conditions for the lane change (e.g., no other vehicle nearby on the neighboring lane). The main results obtained by the authors support the soundness of the FMECA method in identifying and assessing AI-specific failure modes of AI, including adversarial attacks. However, since the method requires significant effort due to state explosion, the authors reinforce that complexity reduction techniques are needed to make the analyses tractable [106].

Nahata et al [107] have crafted a XAI risk prediction model for UGVs based on DTs and RFs, in such a way that the structure of the DTs and the median behavior on RF members would determine the reasoning of the risk prediction engine. The case study presented by Nahata et al [107] aimed to build XAI risk prediction models using as input a dataset including 1,118h of autonomous driving from 20 different UGVs and describe the average driving risk behavior on this dataset. The two main results presented by the authors are as follows: (i.) both the DT and the RF risk prediction models have had appropriate XAI capabilities, since they traced the contribution of each input variable to the risk outcomes throughout the reasoning process, and (ii.) the RF solution has achieved better risk prediction results than single decision trees for the target application.

Groza et al [61] have built an AI-based tool whose aim is to support ophthalmologists in diagnosing two retinal conditions: diabetic retinopathy and age-related macular degeneration. The tool is based on four redundant AI instances: three ML-based ones (namely, ANN, DT, and SVM) for classifying retinal images, and a rule-based system with normative data for retina and a conflict resolution strategy for interleaving the outputs of the ML classifiers. XAI has been deemed relevant by the authors because it is of paramount importance for ophthalmologists to be aware of the reasoning that the diagnosis tool made to reach its final result.

XAI was achieved by Groza et al [61] in two ways: firstly, the DT and the rule-based system are explainable per se; secondly, a rule matrix tool has been used with the ANN and the SVM to translate their complex, black-box models into overapproximate rules of input patterns that were considered to diagnose retinas. The results presented by the authors in case studies aiming to diagnose the state of real retinas, beforehand known to be healthy or unhealthy by experts, corroborate that explainable rules have been obtained with AI along with a degree of confidence on each of these rules [61]. On the other hand, the authors have not explored safety concerns other than explainability, such as the quantitative assessment of the performance and confidence metrics per se.

An application-free tool that supports the white-box analysis of AI-based models is DeepImportance, developed by Gerasimou et al. [108]. Such tool has been crafted with the aim of allowing systematic black-box testing and white-box evaluation of DNNs, including analyses on how their underlying architectures (i.e., specific neurons and interconnects / synapses) affect their overall behavior [108]. With this approach, a better understanding on how each neuron and layer of a DNN contributes with potentially safe and unsafe outputs is facilitated.

The research of Ma et al. [109], in turn, aims to establish a set of quantitative criteria, also based on exploring the underlying architecture of a DNN, to increase the coverage of safety-critical testing – namely, those related to corner cases and adversarial attacks, for which small input perturbations can lead to a significant output shifts [110]. The set of testing criteria, referred to as DeepGauge, has been applied to five DNNs used in image recognition functions and allowed inferring that a greater coverage of critical scenarios has been achieved. Despite the successful results, the authors have

considered that more testing criteria is still needed for an increased test coverage in, e.g., a context of automated test generation for safety-critical systems [109].

DeepXplore has a motivation similar to that of DeepGauge in leveraging test coverage by using data extracted from the internal structure of an AI-based model, and it relies on using AI for that purpose as well. According to Pei et al. [111], experiments have allowed not only detecting that the assessed DNNs failed in dealing with specific types of corner cases, but also in increasing in 3% the accuracy of the models once improvements had been introduced to the design of the DNNs. The authors have not discussed, though, how DeepXplore's AI has been ensured as appropriate for such an application.

Another relevant avenue for building safety arguments based on white box analyses and tests is by using fault injection techniques in such a way that faults are injected on the internal elements of an AI model. By injecting faults to safety-critical AI elements, one can assess how resiliently these are tolerated and whether an unsafe state otherwise undetected in regular tests and analyses can be reached. Two interconnected tools developed by an overlapping group of researchers – namely, TensorFI (TensorFlow Fault Injection) [112] and BinFI (Binary Fault Injection) [113] – are herein highlighted as relevant research on this theme.

TensorFI represents the core fault injection engine for ML-based components of the research by Chen et al. [112], [113]. Even though its development was targeted towards ML implemented with the TensorFlow framework, it is claimed that other frameworks and libraries can also benefit from the underlying fault injection techniques if properly adapted. By assuming the hypothesis that hardware and software faults can be equally represented by corrupting a TensorFlow internal operator, Chen et al. [112] have crafted a tool that allows modeling a wide and plausible set of random and systematic faults that can affect the elements of a computer-based safety-critical system. Experiments with ANNs used in image recognition functions, including those embedded on autonomous driving systems, allowed inferring that TensorFI allows improving the robustness of ML-based elements to faults. Moreover, increasing TensorFI's flexibility to support other frameworks for developing ML, including the C++ version of TensorFlow, are listed as needed improvements [112].

BinFI, in turn, is a binary search-based approach to identify safety-critical ML elements and concentrate the fault injection strategy to these elements instead of performing a mode comprehensive and random fault injection strategy. By means of experiments with DNNs used in autonomous driving systems, Chen et al. [113] have identified that, by using BinFI along with TensorFi, the binary search strategy has outperformed a random fault injection strategy in making ML safer. On the other hand, it has been stressed out that such positive results of the BinFI strategy only apply to ML elements whose error propagation functions are at least approximately monotonic [113].

Finally, Jia et al. [114] have discussed the theoretical relationship between XAI and safety and illustrated their findings by means of a case study in which different types of AI are employed to guide the process of extubation of patients in intensive care units. The main conclusion of the authors is that, even though XAI plays an important role for safety, once it allows tracing back the reasoning performed by the AI in a way that humans are able to understand, XAI is not sufficient for ensuring safety on its own.

Based on the previous discussion, the benefits of XAI and white-box testing come at the expense of the following burdens:

a) There is a higher need for multidisciplinary experts on both safety and AI to support the safety assurance of AI-based systems;
b) XAI on its own poses challenges because it still is an emerging area [52];
c) XAI does not ensure that a system is safe [114];
d) White-box analyses of rather opaque AI models, such as large ANNs and DNNs, are challenging enough to the point of leaning towards unfeasibility on many applications. If their internals could be properly understood during white-box analyses, simpler and more explainable AI models could have been conceived beforehand instead [57], [58]. For instance, among all reviewed research papers of this class, only those by Grushin et al. [105] and Kurd et al. [37] have provided clear, analytic geometry-related XAI capabilities to ANNs.

e: DEFINITION OF SAFETY ASSURANCE PROCESSES SPECIFICALLY FOR AI-BASED SYSTEMS

The last relevant approach for the **safety assurance of AI-based systems** is that it shall be continuously carried out with a **process-oriented approach**, starting on requirements elicitation and extending up to system operation, by monitoring the system outputs with time and comparing them with expected results. It has been argued that such a process-oriented approach shall take into account specific safety assurance techniques for AI [70], [78] and that the typical V-shaped method from non-AI-based safety standards, such as IEC61508 and CENELEC EN50129 [9], [11], is not deemed enough to deal with AI-based systems [115], [116].

In addition to the previous characteristics, extending the safety assurance process of AI-based systems so that it is continuously performed during system operation up to its decommissioning is another difference from non-AI-based systems. This is mostly important for online learning-based systems, since their constant learning changes their architecture as they operate, and the original safety arguments that supported its safety prior to revenue service can be undermined with new, on-demand learned settings.

A landscape of the advantages and disadvantages in crafting a safety assurance process for AI-based systems, based on the further analyses in the present subsection, is presented in Figure 13. Green text boxes indicate potential advantages

of the approach, whereas red text boxes indicate its disadvantages and difficulties.

The previously detailed set of research papers by Kurd et al. [37] and Kurd and Kelly [32], [33], [36] is one of the earliest efforts towards this path as well, since it establishes a systems lifecycle to design explainable ANNs with FSOMs and defines specific design, verification and validation activities at each of its steps. In addition to this set, the research by Douthwaite and Kelly [117], Häring et al. [118], Koopman and Wagner [73], Koopman et al. [119], Mock et al. [120], Pedroza and Adedjouma [116], Pereira and Thomas [121], Salay and Czarnecki [122], and Tarrisse et al. [123] are also relevant examples of systematic, process-oriented means towards the safety assurance of AI-based systems.

Douthwaite and Kelly [117] have crafted a lifecycle for developing safety-critical systems based on Bayesian networks (BNs). For that purpose, five main steps have been identified: (i.) selecting datasets and assessing their applicability; (ii.) creating the BN models per se, including their structure parameterization; (iii.) defining the algorithms to compute BNs; (iv.) selecting supporting modeling frameworks and tools; and (v.) maintaining the system after its deployment. The authors have experimented their workflow on a case study considering an intensive care unit alarm system based on a BN model with 37 random variables and approximately 500 model parameters. Even though the authors have shown promising results with their approach, they highlight that further systematization is still needed, notably on defining failure mode patterns for BNs and tracing safety analysis results to high-level requirements [117].

Koopman and Wagner [73] have presented, without case studies or applications, a framework for the safety validation of UGVs by identifying and discussing factors to increase the robustness of safety arguments whilst balancing the costs of balancing analyses, simulations and tests towards reaching a minimally sufficient robustness. Later, Koopman et al. [119] filled some of these gaps by defining a minimum set of requirements that ML-based systems of UGVs shall observe to reach the foreseen safety goals. It is worth noting that the research by Koopman and Wagner [73] has ultimately led to the crafting of ANSI/UL4600 in 2020 as the first *de jure* standard towards the continuous safety assurance of AI-based systems throughout a system's lifecycle. It is worth noting, though, that ANSI/UL4600 has been conceived with generalization on mind and avoiding at most technology-specific guidelines [15], [124]. Hence, final users of the ANSI/UL4600 standard would still require technological guidance to properly apply it to their products, and developing such guidance would, in turn, still require not only practical expertise, but also an extensive compilation of research efforts still widely pulverized, as shown by means of the present SLR, as well as further advancements on the safety assurance of AI-based systems, given the guidelines for future work discussed about in section VI.

Tarrisse et al. [123] have assessed to what extent the IEC61508:2010 standard [9] could be employed on the safety assurance of AI-based systems and further discussed additional mechanisms that would be needed to make it fully compatible for that application. Based on these analyses, the authors have defined a five-step lifecycle for the safety assurance of systems with AI: (i.) specification; (ii.) data management; (iii.) model development; (iv.) model deployment; and (v.) operation and retirement. The authors have left outside the scope of their research providing further technical remarks on these steps – notably regarding safety assurance techniques –, as well as applying the method on case studies [123].

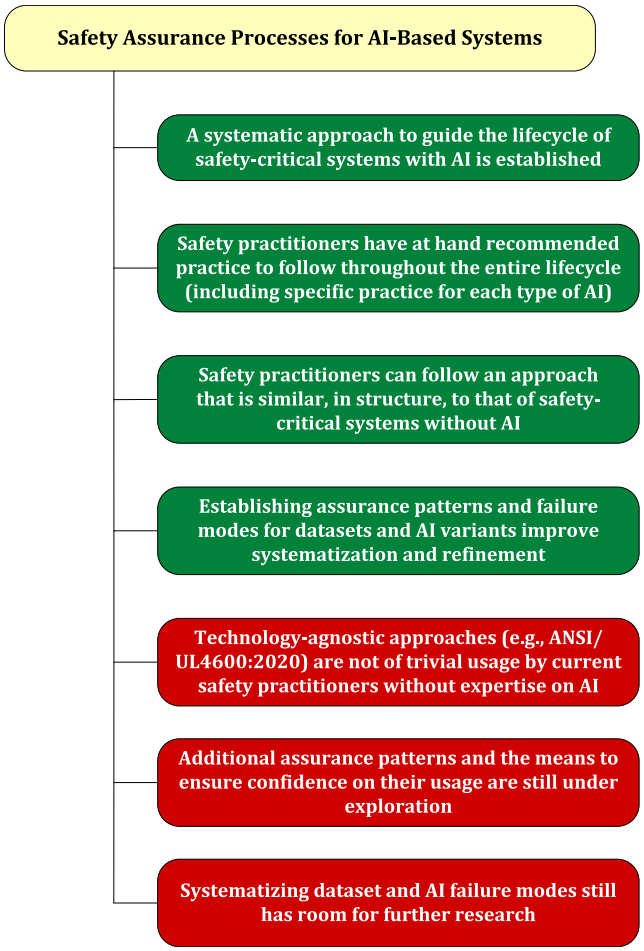

FIGURE 13. Advantages and disadvantages of safety assurance processes for AI-based systems.

Salay and Czarnecki [122] have developed a four-step method to verify safety-critical systems with supervised learning, namely (i.) assessing databases used in the learning process, (ii.) assessing how general the AI response is to input stimuli, (iii.) formally verifying the AI models, and (iv.) repeating such formal verification periodically during the system operation. The authors have provided examples of techniques that could be applied to each of the proposed lifecycle steps, but no case studies on the practical usage of their method have been covered in the research paper [122].

Mock et al [120], in turn, have proposed a 12-step process to define the lifecycle of DNN-based safety-critical

components used within UGV systems. The 12 steps are as follows: (i.) specifying customer-facing functionalities; (ii.) specifying operational design domain context; (iii.) specifying system architecture; (iv.) specifying system functions, notably AI-related ones; (v.) specifying and acquiring training and development data; (vi.) designing ML models; (vii.) pre-processing data; (viii.) training (supervised) ML models; (ix.) post-processing data; (x.) performing tests, verification and validation activities; (xi.) monitoring the system operation; and (xii.) performing maintenance whenever needed. The authors have neither discussed recommended techniques for any of these steps nor presented a practical application of the proposed model by means of, e.g., a case study [120].

Following a similar approach, Häring et al. [118] have developed an 8-step process to guide the lifecycle of AI-based systems – including those with online learning. The eight steps defined by the authors are (i.) context analysis, scope, and aim formulation; (ii.) AI method selection; (iii.) data selection and spotting; (iv.) data preprocessing; (v.) AI model development and training; (vi.) model testing, verification, and validation; (vii.) model application; and (viii.) model modification and updating. Even though the authors have identified that Generative Adversarial Networks (GANs) are useful tools to support the generation of safety-critical scenarios on steps "(i.)" and "(vi.)", the research has limitations similar to those of Mock et al [120] – namely, no deepening of the needed technical activities for each step, and no examples or guidelines for its application [118].

The same trend has also been followed by Pedroza and Adedjouma [116], who have proposed an iterative lifecycle for developing safe-by-design AI-based systems. Each lifecycle iteration includes the following set of 11 steps: (i.) defining missions and goals; (ii.) structuring AI principles; (iii.) performing decompositional analyses; (iv.) structuring AI knowledge bases; (v.) allocating AI techniques; (vi.) selecting knowledge bases; (vii.) designing the detailed AI architecture; (viii.) developing and integrating AI models; (ix.) settling validation benchmarks; (x.) evaluating the AI performance; and (xi.) implementing and deploying the AI system. Safety entwines with this lifecycle by means of situational analyses and the identification of hazards, safety goals, and AI-related malfunctions and faults. Even though the authors have applied the proposed method to build the conceptual design of an autonomous shuttle system in Systems Modeling Language (SysML), no further technical aspects and/or practice have been presented.

The research by Pereira and Thomas [121] also follows a similar approach. On this study, the authors advocate that the lifecycle of an ML-based system shall have at least five steps – namely (i.) requirements specification; (ii.) data management; (iii.) model development; (iv.) model testing and verification; and (v.) model deployment. Furthermore, the authors present a non-exhaustive list of hazards that shall be addressed for safety-critical systems at each of these steps. They also highlight that safety assurance techniques of regular, non-AI-based systems, shall be used along with specific techniques for AI to build sound safety arguments [121].

The conceptual lifecycle and hazards identified by Pereira and Thomas [121] are further exercised on a case study of a self-driving vehicle used in a collaborative human-robot industrial environment. In this case study, the authors illustrate how the lifecycle and the underlying ML hazards can be expanded from a technical standpoint; nevertheless, techniques for assuring that AI is safe are not further explored by the authors.

The SafeML approach proposed by Aslansefat et al. [125] establishes a safety assurance process for classifiers (i.e., supervised learning-based components with discrete outputs). Its safety-related activities range from the selection of appropriate datasets for building and training the classifiers up to the monitoring of the system during its operation. Statistical criteria – notably the cumulative distribution functions of each output class – are used to assess whether safety has been reached with a given confidence. SafeML has been built in such a way to provide proper integration with XAI and security, given their contributions to safety.

Even though the case studies performed by Aslansefat et al. [125] focus on experimental datasets and ML elements not necessarily with a tight link to actual safety-critical applications, as per the SafeML official GitHub project history [126], additional studies have been carried out by other researchers. Bergler [127], for instance, has applied it to an autonomous driving system, focusing especially on the training dataset safety activities. The overall results were positive and supported the soundness of SafeML for ML-based systems used in typical safety-critical applications [127].

An important aspect worth highlighting is that all previously discussed references have included the safety assurance of datasets employed at safety-critical AI design as part of the systems' lifecycle. This is due to the reason that datasets can affect the training and the validation of AI-based systems, leading to, e.g., overfitting and underfitting issues, overly sensitive corner cases and susceptibility to adversarial attacks if they are inappropriate for the target application.

Most research on the safety assurance of datasets used in safety-critical AI-based systems targets to circumvent the aforementioned issues. For instance, Aoki et al. [128] have developed a method to assess labeled datasets used in supervised learning schemes which combines statistical analyses with FTA to assess potential faults on the datasets and exercised it with the recognition of handwritten characters. Boulineau [129] has discussed, among other topics related to safety-critical AI based on supervised learning, a taxonomy of failure modes applicable to labeled datasets and applied it to a train control system which automatically detects track signals. Gauerhof et al. [130] have followed an approach similar to that of Boulineau [129] and defined, among other characteristics on safety-critical AI, means to elicit and assess dataset-related safety requirements. Gauerhof et al. [130] have also explored the practical application of their method on an image dataset applied to obstacle detection by UGVs.

Klaes et al. [131] have discussed the importance of incorporating uncertainty quantification into the safety assurance of AI-based systems, which are tied to the quality of input data and to the underlying architecture and mathematical models of AI by means of a model called Uncertainty Wrapper. Finally, Subbaswamy et al. [132] have presented a framework for analyzing the robustness of ML models to changes on datasets and illustrated its application with a random forest model employed to predict sepsis in hospital patients with different health profiles.

Another theme of interest for systematizing the safety assurance of AI-based systems conceiving safety assurance patterns for specific AI variants, categories, and/or techniques. A safety assurance pattern is a meta-model whose structure defines, for a specific class of systems, a set of safety goals, the contexts in which they are inserted and the arguments that are needed, along with contexts, to fulfill the safety goals [133]. Some research aiming to establish safety assurance patterns includes the efforts by Bragg and Habli [134], Gauerhof et al. [135], and Salay et al. [30].

Bragg and Habli [134] have developed the foundations for an assurance pattern that might be used to support the safety assurance of RL-based systems. The authors have defined that an RL system can be safe on its environment if it satisfies four other lower level goals: (i.) achieving a safe configuration, (ii.) performing a safe reconfiguration when needed, (iii.) transitioning to a fail-safe state when needed, and (iv.) reverting to a safe state when needed. Despite the relevance of assurance patterns as a tool to systematize and simplify the safety assurance of systems, the authors themselves recognized that the lower level goals still need to be further expanded, notably with regard to three main themes: (i.) means to constrain RL for safety, (ii.) means to implement dynamic safety monitoring mechanisms, and (iii.) how to guarantee that online RL ensures its safety on its own [134].

Gauerhof et al. [135] have conceived a safety assurance case for a pedestrian detection function, which is a typical part of an UGV. Even though specific features of the system have been taken into account when building the safety case, such as the usage of convolutional neural networks (CNNs) for image processing, it serves as a relevant pattern not only for other pedestrian detection functions, but also to other object detection features that rely on CNNs. This stems from the fact that the upper goal of the model, defined as '*machine learning function meets all of its safety requirements*', is broad enough for such a generalization.

Salay et al. [30] have proposed a safety case template – hence, an assurance pattern – with a systematic method to generate safety arguments among systems-level and unit-level components of computer vision AI-based systems. The authors have instantiated their template for an object detection task and included a semi-literal solution, which led not only to a qualitative assurance pattern, but also to bound probabilities to reach the applicable safety goals. Even though the research by Salay et al. [30] represents relevant advance on establishing assurance patterns per se, the authors have not provided further details on the means that shall be considered to collect the needed evidence that supports the underlying safety arguments of their assurance pattern [30].

Finally, Cheng et al. [136] have developed an open-source toolbox, called *nn-dependability-kit*, to support the engineering of ANN-based systems used in autonomous driving systems. The foundations of *nn-dependability-kit* are based on an assurance pattern with four major safety goals linked to the AI-based system lifecycle: (i.) ensuring appropriate data collection prior to designing the ANN, (ii.) ensuring proper ANN performance during training and validation, (iii.) ensuring that no potentially unsafe behavior emerges during tests and design generalization, and (iv.) ensuring that no potentially unsafe behavior emerges during actual operation. In order to allow users to reach these goals, specific design and verification techniques (e.g., based on formal methods) are available as part of the toolbox, which has also been positively referred to in previously analyzed studies such as those by Klaes et al. [131] and Gauerhof et al. [135].

4) CONCLUDING REMARKS ON THE STATE OF THE ART OF AI-BASED SYSTEMS

In summary, research correlating safety to AI has significantly evolved since it first emerged in the mid-1980s. The technological evolution of computer systems – notably related to their processing and storage capabilities – have paved the way towards using AI in safety-critical systems and, hence, made the safety assurance of such AI-based systems a major research concern from 2016 onwards.

An overview of the most relevant research towards assuring that AI-based safety-critical systems indeed meet their safety requirements shows that most research on the area spans five major methods towards that objective. These include (i.) black-box testing of AI, (ii.) designing non-AI-based safety envelopes that limit AI response, (iii.) designing fail-safe AI, (iv.) combining explainable AI with white-box analyses, and (v.) establishing a process-oriented approach throughout systems' lifecycle considering specific technical aspects of AI.

Furthermore, the main AI variants that have been exercised follow the current trend of AI research itself, leaning towards machine learning and, more specifically, neural networks, deep learning, and reinforcement learning. Even though this allows safety and AI areas to evolve together, it is deemed that focusing on rather opaque and hard-to-understand models such as deep neural networks is rather challenging for safety, notably because further advancements on simpler and easier-to-understand AI models are still needed.

Finally, with regard to the final results of research papers on the safety assurance of AI-based systems, two main categories have been identified.

The first of them comprises research whose aim is just to propose means to address the safety assurance of AI-based systems. In this case, the methods themselves are the main results presented by the authors, and unless formal

mathematical proof is provided to support the methods' soundness, further research on case studies is usually indicated as the aim for future research. Hence, in these scenarios, one might assume that the research leans towards improving the safety of AI-based systems; nevertheless, there is still no strong conclusion on whether such alleged safety improvements could indeed be reached due to the lack of formal or practical results. This is the case, for instance, of the research by Häring et al. [118], Koopman and Wagner [73], Koopman et al. [119], Mock et al. [120], Pedroza and Adedjouma [116], Salay and Czarnecki [122], Shafaei et al. [76], Tarrisse et al. [123], and Watanabe and Wolf [68].

The second variant includes research in which, along with safety assurance methods, case studies with simulated or real world-based tests are also presented to support the application of the proposed methods. As per the analyses performed throughout subsection ''V-B-3)'', the results presented by the authors are typically positive and supportive of their proposed methods, with a research being proposed for additional improvements. This is the case of Aoki et al. [128], Aslansefat et al [125], Bergler [127], Boulineau [129], Chen et al [100], Chen et al. [112], [113], Cheng et al. [136], Claviere et al. [85], Corso et al. [65], Douthwaite and Kelly [117], Gauerhof et al. [130], [135], Gerasimou et al. [108], Gillula and Tomlin [38], Groza et al [61], Grushin et al. [105], Hussain et al. [70], Jaeger et al. [79], Jia et al. [114], Klaes et al. [131], Kozal and Ksieniewicz [71], Kurd et al. [37], Kurd and Kelly [32], [33], [36], Kuutti et al. [77], Lazarus et al. [58], Ma et al. [109], Mehmood et al. [103], Meltz and Guterman [66], [67], Nahata et al [107], Peng et al. [101], Pereira and Thomas [121], Pei et al. [111], Peruffo et al. [83], Phan et al. [98], Salay et al. [30], Salay et al. [106], Sha et al. [84], Shukla et al. [99], Subbaswamy et al. [132], Wang et al. [56], Wang et al. [86], and Zhao et al. [81]. There are exceptions, though, in which the authors themselves consider that their objectives have not been fully reached, such as Bragg and Habli [134], Lin et al. [80], Machin et al. [75], Sun et al. [69], Tahir and Alexander [21], and Zhao et al. [82].

As a result, one can infer that overall improvements in ensuring safety could be reached in most studies of the second variant. This comes either because the proposed methods themselves have been applied with successful results, or because, even if issues were identified, the authors have discussed relevant future work to circumvent the issues towards allegedly better safety assurance methods or approaches.

VI. NEXT STEPS TOWARDS SAFE AI-BASED SYSTEMS: GUIDELINES FOR FUTURE RESEARCH ON THE SAFETY ASSURANCE OF AI-BASED SYSTEMS

The objective of this paper section is to present an analysis of future work regarding the safety assurance of AI-based safety-critical systems and establish guidelines with relevant research themes yet to be explored in further research towards filling the current gaps on the matter. Since these results stem from the answers to questions Q5 to Q6 (defined in subsection III-E) for all the 329 full-text reviewed C3 references, the guidelines herein presented have a twofold origin. Hence, they not only cover relevant future work identified by the authors of the reviewed research themselves (subsection VI-A), but also those based on the cross-fertilization among the reviewed research and the present research authors' experience with AI and safety-critical systems (subsection VI-B). Finally, the main conclusions of the presented guidelines are covered in subsection VI-C.

A. FIRST PART OF THE GUIDELINES: FUTURE RESEARCH SUGGESTED IN PUBLISHED RESEARCH QUESTION Q5

Out of the 329 C3 papers, 58 of them (17.6%) lack discussion on future work. Hence, the remaining 271 papers in which this topic has been covered served as reference to establish the first part of the guidelines for future work related the safety assurance of AI-based systems.

An overview of the eleven major items that are part of the guidelines for future work as per research recommended on the reviewed references is presented in Figure 14. Further details on each of them, including specific themes and recommended practice stemming from the higher level future work areas, are covered in the following subsections.

1) ADVANCING ON SYSTEMATIC MEANS AND METHODS TO ORIENT THE SAFETY ASSURANCE OF AI-BASED SYSTEMS

The first point of concern for future research is the need to deepen the current efforts towards establishing a process-oriented approach for the safety assurance of AI-based systems during their lifecycle. Such a path has been identified, for instance, by Pedroza and Adedjouma [116] and by Tarrisse et al. [123], who have reinforced that there are few initiatives on the subject [116], most of which still work-in-progress and lacking details [123].

Investing in such future research is considered of paramount importance because, in order to assess whether AI effectively meets the desired safety goals of an application, safety practitioners need beforehand the guidance of means and methods on how to assess safety per se. This a concerning aspect especially because current safety practitioners are not expected to have a deep knowledge of AI, and teaching and training multidisciplinary professionals with expertise in safety and AI is deemed a hard and time-consuming task.

In this sense, establishing a systems-oriented process-based means to deal with the safety assurance of AI-based systems throughout systems' lifecycles and defining an extensive set of 'recommended practice' for each lifecycle step – e.g., focusing on particular techniques for specifying, designing, verifying and validating different AI/ML variants and techniques –, is a relevant direction for future research.

This could be reached, for example, by merging the achieved advancements on safety assurance approaches identified throughout section ''V-B-3)'' – notably, (i.) black-box testing of AI, (ii.) designing non-AI-based safety envelopes

that limit AI response, (iii.) designing fail-safe AI, and (iv.) combining explainable AI to white-box analyses – with the research on 'safety assurance processes for AI-based systems' (subsection ''V-B-3)-e)''). Furthermore, joint efforts along with other future research themes defined in these guidelines would also be of benefit for that.

Special attention shall also be given to the safety assurance of AI-based systems during their operation and maintenance phase. This is relevant especially for systems with online learning, as safety arguments built prior to their operation can become void as systems learn with new data. Potential future work on the assurance of safety-critical systems with online learning require not only assessing the rate with which safety arguments shall be reviewed, but also further advancements on performing automatic safety analyses of AI. Some initial seeds on this subject have been scattered by Cheng and Yan [137] and by Mehmood et al. [103]. Cheng and Yan have reinforced the need for additional research to improve the performance of automated safety analysis tools given the real-time requirements of safety-critical applications [137].

The justification behind advancing on systematic means and methods to the safety assurance of AI-based systems is that, by deepening the definition of activities and recommended practice to avoid and/or mitigate random and systematic faults throughout the lifecycle of safety-critical AI-based systems, safety practitioners would be better prepared to deal with the safety assurance of AI-based systems. With such detailed means and methods at hand, safety practitioners would be able to act in a similar way to process oriented by e.g. IEC61508:2010 [9] and CENELEC EN50129:2018 [11] for non-AI-based systems, circumventing part of the searching and learning efforts practitioners would require to apply a technology-agnostic safety assurance standard for the safety AI-based systems, such as the current version of ANSI/UL4600.

2) DEFINING JUSTIFIABLE AI VARIANTS AND HYPERPAMETERIZATION OF AI MODELS

For safety-critical systems without AI, a system could only be considered safe if, among other conditions, application- specific settings were properly verified and validated as correct and safe for regular and degraded operational modes [138].

With AI, there are two other degrees of freedom when conceiving a solution for a specific application. Firstly, one specific variant of AI shall be chosen among different variants within the same AI type/category (e.g., for discrete supervised learning, ANNs, DT/RFs and SVMs are some of the resources at hand). Secondly, even after a specific AI variant is selected, input data employed while conceiving the models can also influence the internal architecture of the selected AI variant itself by means of the so-called hyperparameters and hyperfunctions. Hence, the choice of AI variants and their hyperparameters and hyperfunctions has a strong potential to influence the behavior of AI, thus directly impacting safety. Wen et al. [25] have raised this specific point of concern on their research.

Hence, future research aiming to explore and improve the current methods of the AI métier in justifying the selection of specific AI types and their settings (i.e., hyperparameters and hyperfunctions) when designing safety-critical systems is relevant. This includes the following topics:

a) Exploring and crafting strategies and techniques to select AI and ML types for safety-critical functions – for instance, by augmenting hazard and risk analyses methods with AI-specific features and failure modes;
b) Formally defining the domains of hyperparameters and hyperfunctions, as well as the strategies to tune them for the target application (e.g., cross-validation schemes, range, scale and step of hyperparameters' variation on each experiment);
c) Making redundant preliminary designs using multiple AI variants and comparing and contrasting them with regard to their performance metrics, distributional shifts, and adversarial attacks within the input datasets. Fault injection mechanisms, such as those implemented in the TensorFI and BinFI tools [112], [113], can be of benefit for this purpose.

3) ASSESSING THE IMPACT OF INPUT DATASETS ON SAFETY

Several researchers have highlighted that datasets employed throughout the design and the operation of safety-critical AI-based systems play a relevant role on the actual safety levels that these systems, as a whole, actually achieve (e.g., Burton et al. [115], Gauerhof et al. [130], Gupta et al. [139], Rajabli et al. [18], Salay and Czarnecki [122], Subbaswamy et al. [132], Watanabe and Wolf [68], Wen et al. [25], Zhang et al. [27]). As mentioned on subsection ''VI-A-2)'', the main reason for this is that datasets themselves influence the architecture of the AI instances crafted for a specific application. Hence, open topics on dataset-related features consist of important themes for further research within the context of AI-based safety-critical systems.

A relevant theme for future research is defining the attributes that a dataset shall possess in order to be deemed adequate for a safety-critical application. Even though high-level foundations for aspects to be observed and avoided are presented in current research (e.g., data bias, dataset shift, concept shift, out-of-domain data [27], quality of labels [140]), general-purpose recommended practice on building and analyzing datasets have not been extensively researched yet.

One point of concern is related to dealing with the representativeness of safety-critical scenarios, which tend to be scarce, in proportion, in datasets which also include records of regular operation of a system. Even though current research indicates that simulation-based approaches are an interesting means to obtain data for safety-critical scenarios without exposing real systems to potentially harmful situations and injecting dataset-related faults (e.g., [25], [87], [115], [139]), the processing of unbalanced databases for

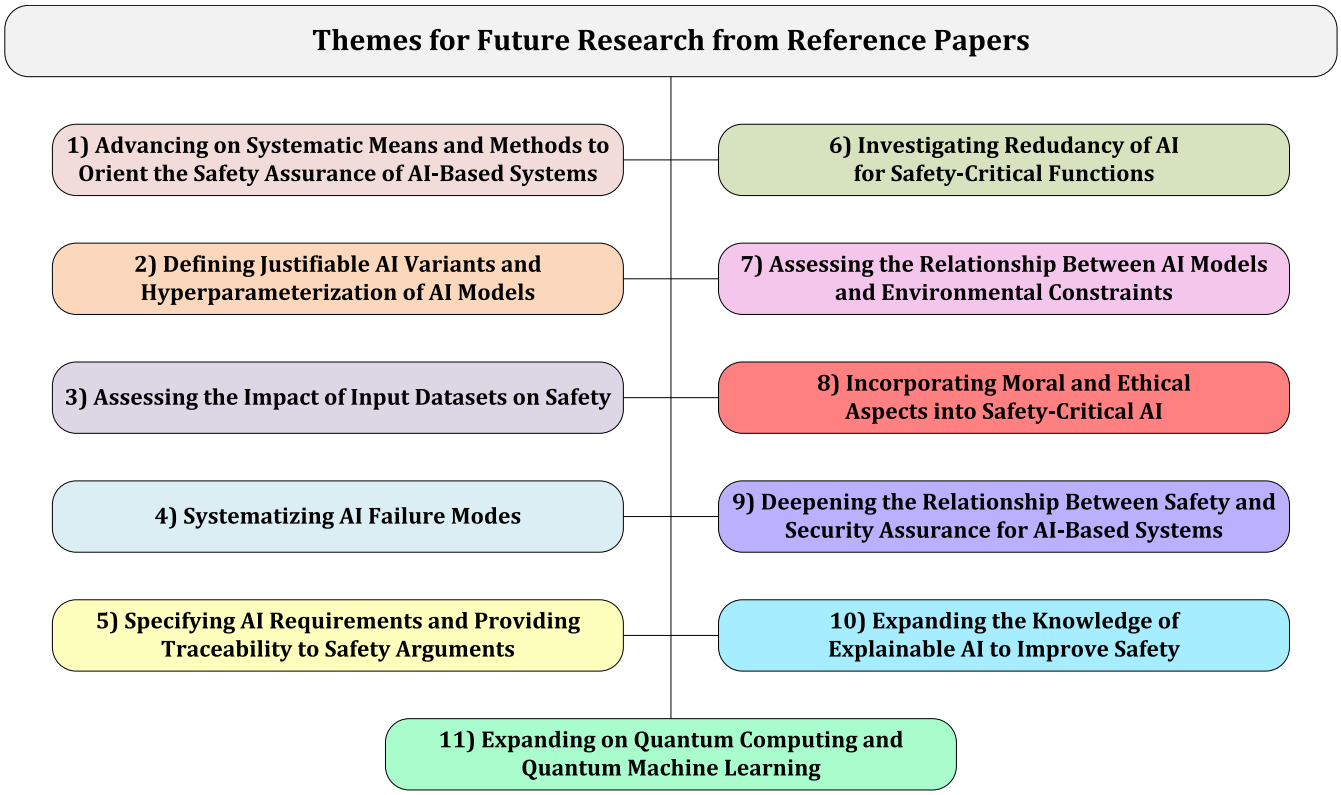

FIGURE 14. Summary of themes for future research on the safety assurance of AI-based systems suggested on reference papers.

safety-critical functions has not been extensively explored. Moreover, the impact of distributional shifts over safety-critical AI, which tends to be greater due to the inherent unbalancing of datasets, has not been explored in depth as well.

The scarcity of data related to safety-critical scenarios also compromises the proper exploration of safety-critical corner cases within datasets. Since corner cases can lead an AI module to present a potentially unsafe behavior with small input perturbations [110], studying them is a relevant concern for safety-critical applications. Future research on this theme shall take into account defining means to identify the representative of corner cases within datasets and methods to assess how AI elements take them into account. Even though general-purpose recommendation can be built to guide that, it is deemed that most efforts shall be application-specific, i.e., that identifying and assessing corner cases depends on the target application per se. Currently, these themes have been explored in more detail only for computer vision (e.g., [135], [139]).

Another important avenue for future work is based on investigating means to quantify and deal with data uncertainty. This theme is relevant because the uncertainties of datasets contents not only reflect their own quality, but also influence the underlying AI architecture and mathematical models [131] – thus impacting on the degree of trust on the outputs produced by such AI. Despite some efforts on that subject by Mjeda and Botterweck [141] and especially the Uncertainty Wrapper by Klaes et al. [131], themes such as the systematic propagation of epistemic uncertainty throughout an AI-based model could still benefit of future research. Moreover, evaluating how to deal with subpar data (e.g., images collected with dirty lenses), data losses, and corrupted data (due to e.g., failure of sensors or adversarial attacks) is also a subset of studies to be considered. The ultimate target of this research area is to ensure that, even in the presence of uncertainties, safety requirements of AI-based systems are met.

Transfer learning, which can be considered as a means to reuse data from one application as initial reference to another [142], is another topic worthy of consideration for further research. So far, Corso and Kochenderfer [142], who have presented a prominent and comprehensive study on the matter, have clearly stated that they still needed further comprehension of transfer learning mechanisms to interpret some results of their results and emphasized that insights on transfer learning algorithms are still needed.

Specifically regarding transfer learning, a first step on future research shall take into account simpler cases in which the datasets employed on the design of AI are sourced from simulated datasets and/or public datasets from the same application domain, but collected on a different environment. Only then further insight on different application domains could be drawn.

Finally, establishing a cost-effective process to ensure safe labeling of datasets for supervised learning solutions also deserves additional investigation on additional research. On the one hand, manual human labeling is bounded to a significantly high failure rate inherent of human beings [140], which might prevent a single labeling chain from being used in safety-critical applications. On the other hand, automated labeling typically relies on semi-supervised or unsupervised learning algorithms, whose safety assurance, in turn, depends on methods which are yet under development, as discussed throughout this paper. Based on such insight, future work aiming to address these limitations is considered welcome to the community.

4) SYSTEMATIZING AI FAILURE MODES

Researchers such as Boulineau [129], Douthwaite and Kelly [117], McDermid et al. [7], and Zhang et al. [27] have highlighted the importance of establishing plausible failure modes for AI models, variants, and techniques.

The main goal of further research on such a theme would be to craft a well-established list of random and systematic AI failure modes, similar in structure and organization to preexisting lists of hardware failure modes, such as the one with CENELEC EN50129:2018 [11]. It is deemed that such a list would feature, for each AI type / variant / approach / technique, an as-exhaustive-as-possible relation of failure modes that could affect it. For instance, the failure modes of an ANN would include potential causes that could change its architecture (e.g., loss of connection between neurons, improper neuron weight), as well as improper hyperparameters (e.g., change on the number of hidden neurons, change on the number of neurons per layer) and improper inputs (e.g., inadequate input dataset).

As already stated in subsection "VI-A-2)", having such a systematic list of AI failure modes would bring benefits in further research on the means to select the most adequate type of AI for an application. Furthermore, it would also allow defining the needed actions to control and mitigate potential safety issues stemming from random and systematic AI failure modes.

5) SPECIFYING AI REQUIREMENTS AND PROVIDING TRACEABILITY TO SAFETY ARGUMENTS

One of the main reasons to consider AI within functions of an engineering system is that specifying these functions is non-trivial in such a way that they cannot be specified neither formally nor exhaustively enough in order to generate a closed-form solution [143]. As a result, the specification of AI-based functions tends to be incomplete, ambiguous and at most partially formal. Such an issue has been discussed and exemplified by several researchers, among which Barzamini et al [144], Boulineau [129], Dey and Lee [15], Koopman and Wagner [73], Kurd et al. [37], and Machin et al. [75].

Moreover, alike every non-formal specification, they are also subject to problems related to the semiotic perception triangle of '*what the system should do*', '*what the system actually does*' and '*what the system is perceived to do*' [7]. These problems can ultimately be translated on systematic failures introduced, consciously or not, at design time.

As a result of this, further research on improving the precision, the exhaustiveness, and the formalism of AI-based safety-critical systems would be of paramount importance towards assuring that AI-based systems are safe. One possible path for that is to provide positive and negative specifications for functions and/or concepts, which clearly state what is within and outside the scope of the said entity.

Another theme worthy of additional research, and which can also benefit from advancements on the aforementioned requirements specification, refers to improving the traceability among system-level requirements and their AI component-specific counterparts. Assuming that better requirements (i.e., more precise, exhaustive and formal) are conceived for AI-related functions, a natural step forward that is to apply the same specification techniques to refine the systems-level requirements into component-level requirements. With such a refinement, the traceability among requirements of different levels becomes less difficult, which has the potential of facilitating the propagation of evidence to build systems-level safety arguments in a bottom-up way (i.e., starting from the low-level safety-critical AI components and moving up the system chain up to its top goals). A starting point towards this is the research by Husen et al. [145], who have briefly explored, among other subjects, AI requirements traceability on a conceptual case study.

6) INVESTIGATING REDUNDANCY OF AI FOR SAFETY-CRITICAL FUNCTIONS

For safety-critical systems without AI, using redundant elements has been considered a feasible approach to meet safety requirements by using 'building blocks' which are not sufficiently safe on their own. Various schemes of redundancy, such as physical redundancy and information redundancy, are also recommended on several standards for safety-critical systems (e.g., [9], [11], [138]).

For AI-based systems, further investigation on whether redundancy is useful practice in leveraging safety is still needed. This concern has been raised by researchers such as Groza et al [61], Kurd et al. [37], and Shafaei et al. [76].

Some correlated topics which are worth analysis in future research involve (i.) assessing the impacts of redundant information in datasets on the robustness of AI instances with regard to safety functions, (ii.) evaluating different schemes of redundant AI elements, and (iii.) assessing potential common-mode failure modes that can affect redundant AI elements.

On (i.), one shall consider that redundant information is not necessarily a replicated representation of the very same data, but instead the presence of multiple different variables which might translate into similar conclusions for the phenomenon of interest. For instance, if one wishes to estimate a person's monthly income, social class and assets net value might be

sufficiently correlated to the point of leading to a converging conclusion.

As per (ii.), a relevant research line comprises conceiving different architectures of redundant AI and comparing and contrasting the results retrieved by their finished designs. The redundant structures might include, for example, several instances of a same AI variant/technique crafted by using different data partitions (e.g., alike *de facto* standardized AI models, such as random forests), or even building ensembles with different AI variants (e.g., building an ANN, an SVM and a DT for the same application). Fault injection tools, such as TensorFI and BinFI [112], [113], can be useful in these activities. Furthermore, mechanisms to build consensus (e.g., majority voting, choice of result with the highest degree of confidence) shall also be further investigated. So far, the study by Groza et al [61] is a starting point for exploring this area.

Finally, (iii.) is directly connected to (i.) and (ii.), since the extent to which common-mode failure modes manifest themselves allow assessing how and to what extent each of the diverse elements of a redundant architecture is affected by the occurrence of the failure mode. It is recommended to perform a case-by-case investigation, depending on the redundancy schemes considered in the assessed safety-critical AI-based systems.

7) ASSESSING THE RELATIONSHIP BETWEEN AI MODELS AND ENVIRONMENTAL CONSTRAINTS

Researchers such as Alexander and Kelly [146], Gauerhof et al. [135], Ruan et al. [147], Tuncali et al. [96] have warned that safety-critical AI-based systems can behave in an unexpected and potentially unsafe way if their design does not take into account the actual conditions and constraints of the environment in which they will effectively operate. As a result, an aspect worthy of concern for future research is to develop means and methods to assess whether environment-related hypotheses are indeed sound for the revenue service of safety-critical systems with AI.

Ensuring that models which represent the operational environment of AI-based components are sufficiently faithful to the real world and that they are coherently applied throughout the system lifecycle becomes a major concern especially because the actual exercise of safety-critical systems on their environments is hardly feasible – especially when dealing with safety-critical scenarios that involve near-misses, as mentioned on subsection ''VI-A-3)''. In these scenarios, simulators, data collected from supposedly similar environments, and transfer learning might be used to fill this gap. Hence, one shall have the means and tools to assess whether these additional elements do not introduce uncertainties and inaccuracies that might undermine the faithfulness and trustworthiness that are necessary for building sound evidence for the safety assurance of AI-based systems.

On reinforcement learning, for instance, these analyses are directly related to the modeling of important hyperparameters and hyperfunctions, such as rewards, the learning rate, and the exploratory test rate. All these features are responsible for balancing exploration and exploitation robustly enough to ensure safety in a dynamic environment.

8) INCORPORATING MORAL AND ETHICAL ASPECTS INTO SAFETY-CRITICAL AI

Burton et al. [148] and Lin and Liu [149] have dealt with a research theme of significant importance especially for fully autonomous safety-critical AI-based systems: the dilemma in which a system, after entering an irreversible state, has to make a decision among a set of alternatives that all lead to undesired, catastrophic outcomes. A hypothetical situation which illustrates such a dilemma could occur, for example, with an UGV that, at some instant of time, is faced with two possible decisions only: colliding at a high speed with the infrastructure, causing the certain death of its single occupant, or colliding at a somewhat lower speed with another vehicle, with the certainty of injuries to the occupants of both vehicles but a lower probability of death for the involved personnel.

The contribution provided by Burton et al. [148] is relevant to deal with such a dilemma, as the authors have identified three gaps – namely, semantic gap, responsibility gap, and liability gap. The consciousness of these gaps allows designers to become aware of potential dilemmas on safety-critical AI-based systems and take the needed action to mitigate them whenever possible and establishing clearer boundaries on when they cannot be avoided and who is to be liable for potentially unsafe scenarios arising from them. The authors provide a twofold recommendation for future research: (i.) the safety assurance process shall be multidisciplinary, involving all potential stakeholders and including, in addition to engineering itself, expert knowledge for law, regulation and governance; and (ii.) providing means for dynamically monitoring and updating safety assurance in order to bridge the underlying gaps of the engineered system.

9) DEEPENING THE RELATIONSHIP BETWEEN SAFETY AND SECURITY ASSURANCE FOR AI-BASED SYSTEMS

Within the context of Industry 4.0, the usage of AI in safety-critical systems has emerged along with significant reliance of engineered systems on fast wireless communication networks [150]. For instance, safety-critical systems of smart cities, along with UGVs running on it, can heavily benefit from the infrastructure of public 5G networks [150]. In this context, a real-time fog computing architecture, in which public processing units can be on-demand requested for data processing, might be used for safety-critical purposes as well [151].

As a result of such a distributed architecture with public networks, a more intricate relationship between security and safety emerges for safety-critical systems that are part of Industry 4.0. Regardless of the usage of AI, assuring proper protection from security attacks is a necessary condition for achieving safety goals,as pointed out by Dey and Lee [15], and Seon and Kim [152].

Specifically when AI is within the loop of safety-critical systems, additional concerns for security shall be considered.

For instance, specific ML variants and models, such as RL, ANNs, and online learning, are susceptible to corner cases in which small perturbations on inputs and/or AI hyperparameters/hyperfunctions can lead to significant changes on outputs [15]. This is significantly worrying because, if an adversarial attack exploiting these corner cases is performed due to a security breach exploited by an intruder, an otherwise safe system can be led to a potentially unsafe state, in which users and the surrounding environment are subject to catastrophic outcomes.

Consequently, future research which aims to deepen the relationship between safety and security for AI-based systems, especially for scenarios in which complying to security requirements is needed for meeting safety requirements, represent important progress for conceiving safety-critical systems with AI. Anastasi et al. [153], for instance, have highlighted the importance of including security analysis techniques within the safety lifecycle of AI-based systems. The usage of graph-based machine learning to leverage security, indicated by Gupta et al. [154] for autonomous vehicles, represents another starting point for that purpose. Furthermore, frameworks for adversarial machine learning, such as the Adversarial Robustness Toolbox [155] and Jespipe [156], are also noteworthy starting points towards improving the resilience of safety-critical AI-based systems to adversarial attacks and their implications.

10) EXPANDING THE KNOWLEDGE OF EXPLAINABLE AI (XAI) TO IMPROVE SAFETY

Based on the information reviewed in subsection "V-B-3)-d)", XAI is still an emerging area per se [52]. As a result, joint efforts in combining XAI with safety still remain a research area with room for significant contribution towards assuring that safety-critical systems with AI meet their safety targets.

Researchers such as Confalonieri et al. [52], Dey and Lee [15], Groza et al [61], Jia et al. [114], Koopman and Wagner [73], Rajabli et al. [18], and Ward and Habli [157] have highlighted the need for additional research on XAI and suggested a trend that includes the following themes:

a) Conceiving guidelines for the structure of arguments generated by XAI in such a way that human practitioners can benefit from XAI (e.g., by improving the link between safety assurance properties and AI interpretability);
b) Assessing desired and necessary features of explainable-by-construction AI, so that sound evidence can be collected for its safety assurance;
c) Deepening the analysis on how to propagate uncertainties throughout the AI reasoning and report them for the AI processing relevant steps;
d) Further investigating means to generate approximate XAI models for hard-to-understand, black-box AI models (e.g., DNNs);
d) Developing systematic evaluation metrics for XAI methods to guide the selection of different types of XAI according to the needs of applications (linked to item "c)", for example);
e) Investigating means to generate safety assurance patterns for XAI.

11) EXPANDING TOWARDS QUANTUM COMPUTING AND QUANTUM MACHINE LEARNING

Incorporating quantum computing into a safety assurance process for AI-based systems is another theme worthy of consideration in future research for two main reasons. Firstly, quantum computing is able to overcome the NP-Hard complexity of the reachability problems in which the formal verification of AI and ML are typically translated into [158]. Secondly, the emergence of quantum machine learning per se can also leverage the conception of new ML models and make their usage feasible in increasingly more intricate safety-critical applications, such as automated medical diagnosis [159] and physics and chemistry processes [160].

Despite the existence of commercial libraries to deal with quantum machine learning [161], its high-scale usage is foreseen as a long-term research rather than to short-to-mid-term ones, though. Current challenges related to building the actual hardware to meet the intended purposes [160], as well as the potentially prohibitive short-term costs for a quantum computing infrastructure for typical consumer-graded applications trending in safety-critical AI research (e.g., transportation and medical support applications) [161] support this.

B. SECOND PART OF THE GUIDELINES: SELF-EXPERIENCE AND CROSS-FERTILIZED FUTURE RESEARCH QUESTION Q6

The second part of the guidelines for future work related to the safety assurance of AI-based systems results from a critical review on each of the C3 publications in order to identify other potential future work. This critical review is not only based on the present SLR authors' experience with safety-critical systems, but also (and especially) on the cross-fertilization among the publications which were reviewed during this SLR.

By means of such critical review, four additional opportunities for future work not discussed within the publications reviewed in this SLR have been identified. They are depicted in Figure 15 and discussed in the following subsections.

1) ASSURING THAT TOOLS AND THIRD-PARTY LIBRARIES FOR DEVELOPING AI-BASED SYSTEMS ARE SAFE

Developing AI typically requires the usage of supporting tools such as simulators and database management systems for tasks such as generating datasets, preprocessing datasets, and testing the behavior of safety-critical AI. In addition to these tools, third-party libraries which implement AI models, such as *scikit-learn* [162], might also be considered for reuse within the design of safety-critical AI.

If the aforementioned tools and third-party libraries have built-in software errors and/or if the underlying hardware in which tools are run fail, improper evidence might be collected, thus leading to misleading arguments on whether the safety-critical AI is indeed safe. For instance, if a software bug on a third-party library remains undetected, its instantiation might be considered safe when it actually is not; similarly, if datasets are corrupted or simulators generate wrong data, the AI itself might be either incorrectly conceived or improperly tested as safe.

As a result, a safety assurance process for AI-based systems shall take into account the risks posed by tools and third-part libraries within systems lifecycle and include activities and techniques meant to capture and mitigate those problems. Future research could, for instance, be based on conceiving a scheme that is similar to that of safety-critical tools present on CENELEC EN50128:2011 [11]. Furthermore, if the tools themselves utilize AI (e.g., dataset labeling using unsupervised or semi-supervised learning), it is recommended to treat them as a safety-critical AI-based system and apply an AI safety assurance method to them.

Similarly, investigating limitations and potential improvements on programming languages typically used in AI design for safety-critical systems is another topic deemed worthy of future research. This item has been raised because most languages conventionally used within the AI area have either not been considered in current standards of safety-critical systems, such as Python and R [9], [138], or are at most weakly recommended, such as C++ and Java [138].

For instance, Wu et al. [74] have provided evidence against using Python for AI-based safety-critical systems, whereas Wozniak et al. [163] have recommended that only strongly typed languages, such as C++, would be recommended for such an application. Since these are still isolated research efforts on this theme, further studies aiming to identify potential safety-related drawbacks of programming languages used in the AI métier and countermeasures to circumvent them could be of benefit to the research community. This research line could ultimately lead to porting and crafting a 'safe subset' of these programming languages, in which only libraries and tools suitable for safety-critical systems would be available.

2) ASSURING SAFETY OF AI-BASED FUNCTIONS ON PROGRAMMABLE HARDWARE DEVICES

On the reviewed literature, no meaningful efforts related to using Programmable Logic Devices (PLDs) to implement safety-critical AI have been identified. Such use for PLDs is deemed plausible for two reasons: firstly, PLDs have been increasingly used in safety-critical embedded systems [164]; secondly, the increasing processing and storage capabilities of modern PLDs, notably Field-Programmable Gate Arrays (FPGAs), make them an interesting alternative for implementing, on their own, dedicated safety-critical AI on embedded systems.

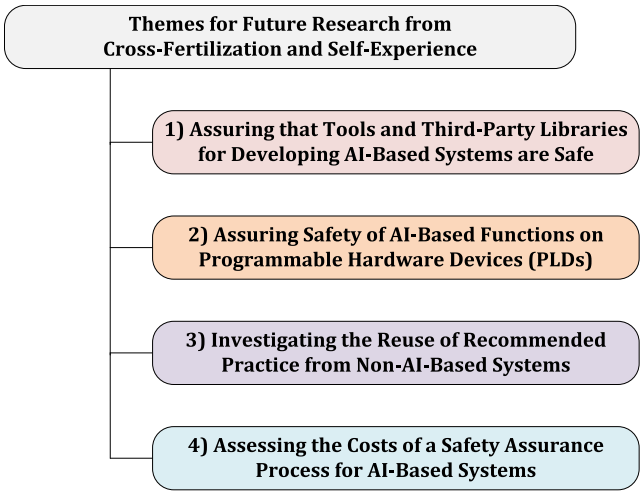

FIGURE 15. Summary of themes for future research on the safety assurance of AI-based systems from cross-fertilization and self-experience.

An important avenue for future research on this theme shall embrace evaluating how to specify, design, implement, verify, validate and ensure the safety of hardware-implemented AI-based safety-critical systems. Special attention shall be paid to how their random and systematic faults can affect the internal architecture of an AI model.

3) INVESTIGATING THE REUSE OF RECOMMENDED PRACTICE FROM NON-AI-BASED SYSTEMS

Researchers such as McDermid et al. [7] and Pereira and Thomas [121] have claimed that safety assurance techniques traditionally applied to non-AI-based systems can be reused for AI-based systems as well. McDermid et al. [7] has deepened this analysis to the point of considering that analyzing hardware-related random faults is among the safety assurance activities which could benefit the most from well-established practice from traditional, non-AI-based systems.

Despite these specific research efforts, no further extensive analyses on the potential of reuse of current safety assurance means, methods and techniques for AI-based systems have been identified. Assuming that potentially reusing these techniques would represent a significant shortcut in preparing current safety practitioners to deal with the safety assurance of AI-based systems, further research on this topic could be of practical benefit. If results supportive of reuse are reached, this could both accelerate and increase the confidence of safety practitioners in working with AI-based systems.

4) ASSESSING THE COSTS OF A SAFETY ASSURANCE PROCESS FOR AI-BASED SYSTEMS

Supposing that several of the other suggested future work is fulfilled and yields a safety assurance process for AI-based systems including a clear set of recommended 'how-to's and practice, it is still needed to investigate whether applying this method is indeed feasible within a cost-constrained environment of real engineering projects.

As a result, the very last item proposed as future work is assessing the costs in applying a process-oriented approach for the safety assurance of AI-based systems in real engineering projects, notably to compare and contrast the needed efforts and technical skills of professionals to current non-AI-based solutions' costs. This could give professionals time and cost estimates for assuring that AI-based systems are safe and, hence, allow companies and professionals to evaluate potential differences not only in the costs for developing and/or buying safety-critical AI, but also in structuring their organization to optimize them to developing and/or using safety-critical systems with AI.

C. CONCLUDING REMARKS ON THE GUIDELINES FOR FUTURE WORK ON THE SAFETY ASSURANCE OF AI-BASED SYSTEMS

The guidelines for future work presented in this section illustrate that, despite the increasing interaction of safety and AI research communities in jointly exploring both areas, as evidenced in this SLR, there is still plenty of room for research on the safety assurance of AI-based system. A set of eleven areas of research were derived from future work suggested on research papers reviewed in this SLR, whereas four additional topics were conceived based on a critical analysis carried out by combining the cross-fertilization of the reviewed research papers with the expertise of the authors of this SLR in AI and safety-critical systems.

It is deemed that, among all the raised topics for future work, establishing a process-oriented method for the safety assurance of AI-based systems is a natural first step. The reasoning that backs this recommendation up is that, alike with traditional safety-critical systems lacking AI, the training of safety practitioners with expertise for AI-based systems is facilitated once there is a systematic approach for dealing the safety lifecycle, along with the needed safety activities and recommended practice for each of its steps.

An outlook for future research on this area is to use pre-existing safety assurance processes for AI-based systems – established on, e.g., the AI technology-agnostic ANSI/UL4600:2020 standard and the research papers discussed in subsection "V-B-3)-e)" – as templates and improve them in a twofold way. Firstly, it is pertinent to make sure that gaps on the safety lifecycles of the 'template' processes are filled with steps and activities that cover the missing relevant AI-related safety-critical themes they lack. Secondly, it is worth compiling preexisting safety assurance techniques which have not been contextualized within the 'template' processes (e.g., techniques cited in subsections "V-B-3)-a)" to "V-B-3)-d)") and map them onto the safety lifecycle steps along with how their usage would be recommended. This could be achieved by departing from the positive and negative outcomes of these techniques, as per reported in the subsections "V-B-3)-a)" to "V-B-3)-d)" of this paper and the research therein quoted.

For instance, if further research on refining the simplex architecture defined in the subsection "V-B-3)-c)" is performed, it is worth considering that an approach similar to that of Mehmood et al. [103] might be unfeasible for, e.g., practical safety-critical embedded systems due to its stringent storage requirements. As a result, expanding correlated research shall start by learning with the limitations of existing solutions and ultimately trying either to optimize them or to follow a different approach if such an optimization is deemed unfeasible or unjustifiable.

Moreover, other avenues on future research involve exploring specific techniques to be used within steps of the aforementioned systems lifecycle. These specific technical topics include (i.) tightening the justification of AI hyperparameterization, (ii.) analyzing the adequacy of datasets, (iii.) systematizing AI failure modes, (iv.) improving the specification of AI, (v.) exploring AI redundancy, (vi.) tightening assumptions and improving models of the environment at which the AI is used, (vii.) dealing with moral and ethical aspects, (viii.) deepening the relationship between safety and security, (ix.) exploring explainable AI to improve safety, (x.) expanding on quantum computing and quantum machine learning, (xi.) assuring safety of PLD-implemented AI, (xii.) assuring safety of AI development tools, (xiii.) reusing practice to assure safety of non-AI-based systems, and (xiv.) estimating costs for assuring AI-based systems.

Specifically for items (i.) to (x.), a starting point for incorporating them on future research should also take into consideration the positive and negative results of their prior experimentation on preexisting research, quoted throughout subsections "VI-A-2)" to "VI-A-11)". For instance, when dealing with transfer learning as part of item '(ii.) analyzing the adequacy of datasets', future research can be initially tightened to the challenges posed on the state of the art – namely, a better understanding of transfer learning mechanisms [142] –, and just then widened to other relevant related areas (e.g., investigating different transfer learning applications on safety-critical systems and developing assurance patterns for transfer learning).

For items (xi.) to (xiv.), in turn, the lack of consistent previous research on them makes it harder to constrain the starting points of such themes to potentially promising paths. In these scenarios, the general guidance provided in the subsections "VI-B-1)" to "VI-B-4)" is recommended for that purpose.

Finally, it is considered that future research could also benefit from a tighter integration between AI and safety researchers, since they are related to adapting typical practice of the AI field to stricter requirements imposed by safety-critical applications. This integration is also important because it is envisioned that safety professionals shall excel in multidisciplinary knowledge in both safety and AI in order to coordinate and perform the needed tasks to ensuring that these systems meet their safety requirements.

VII. CONCLUSION

The objective of this paper was to present an overview on the state of the art and guidelines for future research on the safety assurance of AI-based systems by means of an SLR comprising texts published until August 26th, 2022. As justified in

section II, the main contribution of this research is not only to go beyond the scope of other SLRs by covering a broader range of applications and following a well-controlled and reproducible process on peer-reviewed publications only, but also to present an updated landscape on the safety assurance of systems with AI and introduce guidelines for relevant future work. The latter has been reached through a critical analysis of the reviewed references stemming from both the cross-fertilization among the reviewed references and the own experience of the SLR authors on safety assurance and AI.

The six-step SLR, carried out as per section III and leading to the results presented in section IV, covered a total of 5090 references, among which a subset of 329 publications which somehow directly address the safety assurance of AI-based systems was considered in further steps. By means of these 329 publications, it has been concluded that research on the theme has sharply increased especially over the last years (2016 onwards) and that increasingly more research is also expected for the forthcoming years.

Based on the detailed review of the aforementioned 329 publications subset, it has been identified that the safety assurance of AI-based systems has been carried out following five main approaches to build safety arguments: (i.) performing exhaustive black-box testing of AI, (ii.) constraining the response of safety-critical AI by means of a non-AI-dependent safety envelope, (iii.) designing fail-safe AI, (iv.) combining explainable AI with its white-box analyses, and (v.) establishing a continuous, process-oriented safety assurance process throughout systems' lifecycles. The overall conclusion is that current research on the safety assurance of AI-based systems indicates significant improvements towards allowing AI to be used and proven as safe; nevertheless, further advancements are still needed to fully reach this result. Details on each of the aforementioned safety assurance approaches, including their pros, cons, state of the art and current limitations, were explored in section V.

Guidelines for potential future research topics have also been presented in this research. These include not only recurrent themes indicated by other researchers, but also additional topics which stemmed from both the cross-fertilization of the reviewed references and the experience of the authors of this SLR with AI and safety. These guidelines are presented in two parts on section VI. Among all its items, two main conclusions are highlighted. The first of them is the need for a better integration of AI and safety métiers, so that the resulting methods and approaches for the safety assurance of AI-based systems can be combined in an effective way. It is expected that this research aids paving this way.

The second highlight of the guidelines is the need for further research towards a systematic, process-oriented approach for the safety assurance of AI-based systems which includes recommended technical guidelines to deal with AI-specific aspects. These include, but are not limited to, improving the breadth and the depth of preexisting safety assurance processes, analyzing datasets, eliciting AI requirements, choosing and adjusting AI hyperparameters, defining the desired features of safety-critical explainable AI (e.g., uncertainty propagation), systematizing AI failure modes, and defining means to design, verify and validate safety-critical AI-based functions implemented on PLDs. The basic strategy recommended to deal with these themes is to consider the positive and negative results of preexisting research as starting points to guide research efforts, and then broaden the scope of such research to other themes once the known gaps have been either filled or discarded. It is ultimately expected that, by following the herein defined guidelines, safety practitioners are provided with a safety assurance experience closer to that of non-AI-related standards (e.g., IEC61508-derived ones) than the technology-agnostic approach of the 2020 *de jure* ANSI/UL4600 standard for safety-critical AI-based systems.

Finally, it is worth mentioning that the results herein reported are the first step of the authors' research. The next envisioned step is to contribute with the filling of the gaps identified in the guidelines for future work by establishing a safety assurance process-oriented approach including a set of recommended techniques for each of its steps and a detailed workflow to guide its application. The ultimate objective of the research is to apply this safety assurance method to safety-critical AI-based systems and evaluate its convergence towards results that ensure that a system conceived with it is indeed safe.

SUPPLEMENTARY MATERIAL

Additional results of the SLR, including the formal definition of the SLR Search Language, further justification on its expressions, its instantiations for each search engine, and the full list of reviewed references along with their analyses (e.g., attribution of categories C1-C5, attribution of Q-index, answers to questions Q1-Q6 and additional bibliometrics), are available within the technical report [165]. This report has been made public on Zenodo.org not only as an open science effort, but also as a means to increase the transparency of the research and support the findings reported in this paper.

REFERENCES

[1] Z. Allam and Z. A. Dhunny, "On big data, artificial intelligence and smart cities," *Cities*, vol. 89, pp. 80–91, Jun. 2019, doi: 10.1016/j.cities.2019.01.032.

[2] M. Xu and H. Liu, "A flexible deep learning-aware framework for travel time prediction considering traffic event," *Eng. Appl. Artif. Intell.*, vol. 106, Nov. 2021, Art. no. 104491, doi: 10.1016/J.ENGAPPAI.2021.104491.

[3] C. Liu, W. B. Rouse, and D. Belanger, "Understanding risks and opportunities of autonomous vehicle technology adoption through systems dynamic scenario modeling—The American insurance industry," *IEEE Syst. J.*, vol. 14, no. 1, pp. 1365–1374, Mar. 2020, doi: 10.1109/JSYST.2019.2913647.

[4] EASA. (2020). *Artificial Intelligence Roadmap—A Human-Centric Approach to AI in Aviation*. Accessed: Nov. 23, 2022. [Online]. Available: https://www.easa.europa.eu/en/downloads/109668/en

[5] J. Doppelbauer. (2018). *Command and Control 4.0*. IRSE News. Accessed: Nov. 23, 2022. [Online]. Available: https://www.era.europa.eu/system/files/2022-10/Command%20and%20Control%204.0.pdf

[6] I. Allende, N. M. Guire, J. Perez-Cerrolaza, L. G. Monsalve, J. Petersohn, and R. Obermaisser, "Statistical test coverage for linux-based next-generation autonomous safety-related systems," *IEEE Access*, vol. 9, pp. 106065–106078, 2021, doi: 10.1109/ACCESS.2021.3100125.
[7] J. McDermid, Y. Jia, and I. Habli, "Towards a framework for safety assurance of autonomous systems," in *Proc. CEUR Workshop*, vol. 2419, 2019, pp. 1–7.
[8] I. Habli, T. Lawton, and Z. Porter, "Artificial intelligence in health care: Accountability and safety," *Bull. World Health Org.*, vol. 98, no. 4, pp. 251–256, Apr. 2020, doi: 10.2471/BLT.19.237487.
[9] *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems (7 Parts)*, document IEC, ISO/IEC61508:2010, Geneve, Switzerland, 2010.
[10] *Railway Applications—The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS)—Part 1: Generic RAMS Process*, document CENELEC, EN50126-1:2017, Brussels, Belgium, 2017.
[11] *Railway Applications Communication, Signalling and Processing Systems Safety-Related Electronic Systems for Signalling*, document CENELEC, EN50129:2018, Brussels, Belgium, 2018.
[12] *Design Assurance Guidance for Airborne Electronic Hardware*, document RTCA, DO-254, Washington, DC, USA, 2000.
[13] S. Ballingall, M. Sarvi, and P. Sweatman, "Safety assurance concepts for automated driving systems," *SAE Int. J. Adv. Current Practices Mobility*, vol. 2, no. 3, pp. 1528–1537, 2020, doi: 10.4271/2020-01-0727.
[14] W. M. D. Chia, S. L. Keoh, C. Goh, and C. Johnson, "Risk assessment methodologies for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 16923–16939, Oct. 2022, doi: 10.1109/TITS.2022.3163747.
[15] S. Dey and S.-W. Lee, "Multilayered review of safety approaches for machine learning-based systems in the days of AI," *J. Syst. Softw.*, vol. 176, Jun. 2021, Art. no. 110941, doi: 10.1016/j.jss.2021.110941.
[16] S. Kabir, "An overview of fault tree analysis and its application in model based dependability analysis," *Expert Syst. Appl.*, vol. 77, pp. 114–135, Jul. 2017, doi: 10.1016/j.eswa.2017.01.058.
[17] A. M. Nascimento, "A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 12, pp. 4928–4946, Dec. 2020, doi: 10.1109/tits.2019.2949915.
[18] N. Rajabli, F. Flammini, R. Nardone, and V. Vittorini, "Software verification and validation of safe autonomous cars: A systematic literature review," *IEEE Access*, vol. 9, pp. 4797–4819, 2021, doi: 10.1109/ACCESS.2020.3048047.
[19] A. Rawson and M. Brito, "A survey of the opportunities and challenges of supervised machine learning in maritime risk analysis," *Transp. Rev.*, vol. 43, no. 1, pp. 108–130, Jan. 2023, doi: 10.1080/01441647.2022.2036864.
[20] G. Siedel, S. Voß, and S. Vock, "An overview of the research landscape in the field of safe machine learning," in *Proc. Saf. Eng., Risk, Rel. Anal., Res. Posters*, vol. 13, Nov. 2021, Art. no. V013T14A045, doi: 10.1115/IMECE2021-69390.
[21] Z. Tahir and R. Alexander, "Coverage based testing for V&V and safety assurance of self-driving autonomous vehicles: A systematic literature review," in *Proc. IEEE Int. Conf. Artif. Intell. Test. (AITest)*, Aug. 2020, pp. 23–30, doi: 10.1109/AITEST49225.2020.00011.
[22] 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, pp. 1–74, Apr. 2022, doi: 10.1007/S10515-022-00337-X.
[23] Y. Wang and M. P. Chapman, "Risk-averse autonomous systems: A brief history and recent developments from the perspective of optimal control," *Artif. Intell.*, vol. 311, Oct. 2022, Art. no. 103743, doi: 10.1016/j.artint.2022.103743.
[24] 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, doi: 10.1108/IMDS-07-2021-0419.
[25] H. Wen, F. Khan, M. T. Amin, and S. Z. Halim, "Myths and misconceptions of data-driven methods: Applications to process safety analysis," *Comput. Chem. Eng.*, vol. 158, Feb. 2022, Art. no. 107639, doi: 10.1016/j.compchemeng.2021.107639.
[26] J. Zhang and J. Li, "Testing and verification of neural-network-based safety-critical control software: A systematic literature review," *Inf. Softw. Technol.*, vol. 123, Jul. 2020, Art. no. 106296, doi: 10.1016/j.infsof.2020.106296.
[27] X. Zhang, F. T. S. Chan, C. Yan, and I. Bose, "Towards risk-aware artificial intelligence and machine learning systems: An overview," *Decis. Support Syst.*, vol. 159, Aug. 2022, Art. no. 113800, doi: 10.1016/j.dss.2022.113800.
[28] S. A. Asadollah, D. Sundmark, S. Eldh, H. Hansson, and W. Afzal, "10 years of research on debugging concurrent and multicore software: A systematic mapping study," *Softw. Quality J.*, vol. 25, no. 1, pp. 49–82, Jan. 2016, doi: 10.1007/S11219-015-9301-7.
[29] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proc. EASE 12th Int. Conf. Eval. Assessment Softw. Eng.*, Jun. 2008, pp. 68–77, doi: 10.14236/ewic/EASE2008.8.
[30] R. Salay, K. Czarnecki, H. Kuwajima, H. Yasuoka, V. Abdelzad, C. Huang, M. Kahn, V. D. Nguyen, and T. Nakae, "The missing link: Developing a safety case for perception components in automated driving," in *Proc. SAE Tech. Paper Ser.*, Mar. 2022, pp. 1–13, doi: 10.4271/2022-01-0818.
[31] Elsevier B.V. (2022). *What is the Difference Between ScienceDirect and Scopus Data? Data as a Service Support Center*. Accessed: Sep. 30, 2022. [Online]. Available: https://service.elsevier.com/app/answers/detail/a_id/28240/supporthub/dataasaservice/p/17729/
[32] Z. Kurd and T. P. Kelly, "Using fuzzy self-organising maps for safety critical systems," in *Computer Safety, Reliability, and Security* (Lecture Notes in Computer Science), vol. 3219. Berlin, Germany: Springer, 2004, pp. 17–30.
[33] Z. Kurd and T. P. Kelly, "Using safety critical artificial neural networks in gas turbine aero-engine control," in *Proc. 24th Int. Conf. Comput. Saf., Rel., Secur. (SAFECOMP)*, vol. 3688, 2005, pp. 136–150, 2005, doi: 10.1007/11563228_11.
[34] B. Cukic, E. Fuller, M. Mladenovski, and S. Yerramalla, "Run-time assessment of neural network control systems," in *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*. New York, NY, USA: Springer, 2006, pp. 257–269, doi: 10.1007/0-387-29485-6_10.
[35] L. Pullum and B. J. Taylor, "Risk and hazard analysis for neural network systems," in *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*. New York, NY, USA: Springer, 2006, pp. 33–49, doi: 10.1007/0-387-29485-6_3.
[36] Z. Kurd and T. P. Kelly, "Using fuzzy self-organising maps for safety critical systems," *Rel. Eng. Syst. Saf.*, vol. 92, no. 11, pp. 1563–1583, Nov. 2007, doi: 10.1016/j.ress.2006.10.005.
[37] Z. Kurd, T. Kelly, and J. Austin, "Developing artificial neural networks for safety critical systems," *Neural Comput. Appl.*, vol. 16, no. 1, pp. 11–19, Oct. 2006, doi: 10.1007/s00521-006-0039-9.
[38] J. H. Gillula and C. J. Tomlin, "Guaranteed safe online learning via reachability: Tracking a ground target using a quadrotor," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 2723–2730, doi: 10.1109/ICRA.2012.6225136.
[39] G. Mancini, "Collection, processing and use of data," *Nucl. Eng. Des.*, vol. 93, nos. 2–3, pp. 181–186, 1986, doi: 10.1016/0029-5493(86)90217-7.
[40] T. Washio, M. Kitamura, K. Kotajima, and K. Sugiyama, "Automated generation of nuclear power plant safety information," in *Proc. Power Plant Dyn., Control Test. Symp.*, vol. 1, 1986, pp. 39.01–39.17.
[41] F. L. Cho, "Expert system application in equipment risk assessment for nuclear power plants," in *Proc. Comput.-Aided Eng. Appl. Pressure Vessels Piping Conf.*, vol. 126, 1987, pp. 27–32.
[42] R. C. Erdmann and B. K.-H. Sun, "Expert system approach for safety diagnosis," *Nucl. Technol.*, vol. 82, no. 2, pp. 162–172, Aug. 1988, doi: 10.13182/NT88-A34105.
[43] R. E. Uhrig, "Use of probabilistic risk assessment (PRA) in expert systems to advise nuclear plant operators and managers," *Proc. SPIE*, vol. 937, pp. 210–215, Mar. 1988, doi: 10.1117/12.946977.
[44] B. Frisch, J. Lecinena, C. Preyssl, A. Saleem, F. Stolle, and E. Tosini, "ERES—An expert system for ESA risk assessment and management," *Sci. Technol. Ser.*, vol. 93, pp. 161–171, Jan. 1997.
[45] R. Vaidhyanathan and V. Venkatasubramanian, "Experience with an expert system for automated HAZOP analysis," *Comput. Chem. Eng.*, vol. 20, pp. S1589–S1594, Jan. 1996, doi: 10.1016/0098-1354(96)00270-0.

[46] J. Fox, "Expert systems for safety-critical applications: Theory, technology and applications," in *Proc. IEE Colloq. Knowl.-Based Syst. Saf. Crit. Appl.*, no. 109, May 1994, pp. 5-1–5-5.

[47] M. Kitamura, "Knowledge engineering approach to risk management and decision-making problems," *Rel. Eng. Syst. Saf.*, vol. 38, nos. 1–2, pp. 67–70, Jan. 1992.

[48] G. Xie, D. Xue, and S. Xi, "TREE-EXPERT: A tree-based expert system for fault tree construction," *Reliab. Eng. Syst. Saf.*, vol. 40, no. 3, pp. 295–309, 1993, doi: 10.1016/0951-8320(93)90066-8.

[49] E. A. Averbukh, "Neural network models and statistical tests as flexible base for intelligent fault diagnosis," *Annu. Rev. Autom. Program.*, vol. 17, no. 10, pp. 259–266, 1992, doi: 10.1016/S0066-4138(09)91043-6.

[50] X. Z. Wang, B. H. Chen, S. H. Yang, and C. McGreavy, "Neural nets, fuzzy sets and digraphs in safety and operability studies of refinery reaction processes," *Chem. Eng. Sci.*, vol. 51, no. 10, pp. 2169–2178, 1996, doi: 10.1016/0009-2509(96)00074-7.

[51] C.-H. Wei, "Developing freeway lane-changing support systems using artificial neural networks," *J. Adv. Transp.*, vol. 35, no. 1, pp. 47–65, Sep. 2001, doi: 10.1002/atr.5670350105.

[52] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 1, Jan. 2021, Art. no. e1391, doi: 10.1002/WIDM.1391.

[53] A. Miller, "The intrinsically linked future for human and artificial intelligence interaction," *J. Big Data*, vol. 6, no. 1, pp. 1–9, May 2019, doi: 10.1186/S40537-019-0202-7.

[54] D. Liu, H. Kong, X. Luo, W. Liu, and R. Subramaniam, "Bringing AI to edge: From deep learning's perspective," *Neurocomputing*, vol. 485, pp. 297–320, May 2022, doi: 10.1016/J.NEUCOM.2021.04.141.

[55] G. Giray, "A software engineering perspective on engineering machine learning systems: State of the art and challenges," *J. Syst. Softw.*, vol. 180, Oct. 2021, Art. no. 111031, doi: 10.1016/J.JSS.2021.111031.

[56] Q. Wang, G. Kou, L. Chen, Y. He, W. Cao, and G. Pu, "Runtime assurance of learning-based lane changing control for autonomous driving vehicles," *J. Circuits, Syst. Comput.*, vol. 31, no. 14, Sep. 2022, Art. no. 2250249, doi: 10.1142/S0218126622502498.

[57] F. Flammini, S. Marrone, R. Nardone, M. Caporuscio, and M. D'Angelo, "Safety integrity through self-adaptation for multi-sensor event detection: Methodology and case-study," *Future Gener. Comput. Syst.*, vol. 112, pp. 965–981, Nov. 2020, doi: 10.1016/j.future.2020.06.036.

[58] C. Lazarus, J. G. Lopez, and M. J. Kochenderfer, "Runtime safety assurance using reinforcement learning," in *Proc. AIAA/IEEE 39th Digit. Avionics Syst. Conf. (DASC)*, Oct. 2020, pp. 1–9, doi: 10.1109/DASC50938.2020.9256446.

[59] Y. Chandak, S. M. Jordan, G. Theocharous, M. White, and P. S. Thomas, "Towards safe policy improvement for non-stationary MDPs," in *Proc. 34th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2020, pp. 9156–9168. Accessed: Nov. 23, 2022. [Online]. Available: https://dl.acm.org/doi/10.5555/3495724.3496492

[60] J. Hernández-Orallo, F. Martínez-Plumed, S. Avin, J. Whittlestone, and S. Ó. Héigeartaigh, "AI paradigms and AI safety: Mapping artefacts and techniques to safety issues," in *Proc. 24th Eur. Conf. Artif. Intell., (ECAI) Including 10th Conf. Prestigious Appl. Artif. Intell. (PAIS)*, vol. 325, 2020, pp. 2521–2528, doi: 10.3233/FAIA200386.

[61] A. Groza, L. Toderean, G. A. Muntean, and S. D. Nicoara, "Agents that argue and explain classifications of retinal conditions," *J. Med. Biol. Eng.*, vol. 41, pp. 730–741, Sep. 2021, doi: 10.1007/s40846-021-00647-7.

[62] I. Ruchkin, M. Cleaveland, R. Ivanov, P. Lu, T. Carpenter, O. Sokolsky, and I. Lee, "Confidence composition for monitors of verification assumptions," in *Proc. ACM/IEEE 13th Int. Conf. Cyber-Phys. Syst. (ICCPS)*, May 2022, pp. 1–12, doi: 10.1109/ICCPS54341.2022.00007.

[63] P. Musau, N. Hamilton, D. M. Lopez, P. Robinette, and T. T. Johnson, "On using real-time reachability for the safety assurance of machine learning controllers," in *Proc. IEEE Int. Conf. Assured Autonomy (ICAA)*, Mar. 2022, pp. 1–10, doi: 10.1109/ICAA52185.2022.00010.

[64] Y. Bai, Z. Huang, H. Lam, and D. Zhao, "Rare-event simulation for neural network and random forest predictors," *ACM Trans. Model. Comput. Simul.*, vol. 32, no. 3, pp. 1–33, Jul. 2022, doi: 10.1145/3519385.

[65] A. Corso, R. Moss, M. Koren, R. Lee, and M. Kochenderfer, "A survey of algorithms for black-box safety validation of cyber-physical systems," *J. Artif. Intell. Res.*, vol. 72, pp. 377–428, Oct. 2021, doi: 10.1613/JAIR.1.12716.

[66] D. Meltz and H. Guterman, "RobIL—Israeli program for research and development of autonomous UGV: Performance evaluation methodology," in *Proc. IEEE Int. Conf. Sci. Electr. Eng. (ICSEE)*, Nov. 2016, pp. 1–5, doi: 10.1109/ICSEE.2016.7806157.

[67] D. Meltz and H. Guterman, "Functional safety verification for autonomous UGVs—Methodology presentation and implementation on a full-scale system," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 3, pp. 472–485, Sep. 2019, doi: 10.1109/TIV.2019.2919460.

[68] T. Watanabe and D. Wolf, "Verisimilar percept sequences tests for autonomous driving intelligent agent assessment," in *Proc. Latin Amer. Robotic Symp., Brazilian Symp. Robot. (SBR) Workshop Robot. Educ. (WRE)*, Nov. 2018, pp. 188–193, doi: 10.1109/LARS/SBR/WRE.2018.00048.

[69] J. Sun, H. Zhou, H. Xi, H. Zhang, and Y. Tian, "Adaptive design of experiments for safety evaluation of automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14497–14508, Sep. 2022, doi: 10.1109/TITS.2021.3130040.

[70] M. Hussain, N. Ali, and J.-E. Hong, "DeepGuard: A framework for safeguarding autonomous driving systems from inconsistent behaviour," *Automated Softw. Eng.*, vol. 29, no. 1, pp. 1–32, May 2022, doi: 10.1007/s10515-021-00310-0.

[71] J. Kozal and P. Ksieniewicz, "Imbalance reduction techniques applied to ECG classification problem," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.*, vol. 11872, 2019, pp. 323–331, doi: 10.1007/978-3-030-33617-2_33.

[72] C. Harper and P. Caleb-Solly, "Towards an ontological framework for environmental survey hazard analysis of autonomous systems," in *Proc. CEUR Workshop*, vol. 2808, 2021, pp. 1–7.

[73] P. Koopman and M. Wagner, "Toward a framework for highly automated vehicle safety validation," in *Proc. SAE Tech. Paper Ser.*, Apr. 2018, pp. 1–13, doi: 10.4271/2018-01-1071.

[74] H. Wu, D. Lv, T. Cui, G. Hou, M. Watanabe, and W. Kong, "SDLV: Verification of steering angle safety for self-driving cars," *Formal Aspects Comput.*, vol. 33, no. 3, pp. 325–341, Jun. 2021, doi: 10.1007/s00165-021-00539-2.

[75] M. Machin, J. Guiochet, H. Waeselynck, J.-P. Blanquart, M. Roy, and L. Masson, "SMOF: A safety monitoring framework for autonomous systems," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 5, pp. 702–715, May 2018, doi: 10.1109/TSMC.2016.2633291.

[76] S. Shafaei, S. Kugele, M. H. Osman, and A. Knoll, "Uncertainty in machine learning: A safety perspective on autonomous driving," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 11094, 2018, pp. 458–464, doi: 10.1007/978-3-319-99229-7_39.

[77] S. Kuutti, R. Bowden, H. Joshi, R. de Temple, and S. Fallah, "Safe deep neural network-driven autonomous vehicles using software safety cages," in *Proc. 20th Int. Conf. Intell. Data Eng. Automated Learn., (IDEAL)*, vol. 11872, 2019, pp. 150–160, doi: 10.1007/978-3-030-33617-2_17.

[78] S. Schirmer, C. Torens, F. Nikodem, and J. Dauer, "Considerations of artificial intelligence safety engineering for unmanned aircraft," in *Proc. Workshops, ASSURE, DECSoS, SASSUR, STRIVE, WAISE Co-Located With 37th Int. Conf. Comput. Saf., Rel. Secur., (SAFECOMP)*, vol. 11094. Berlin, Germany: Springer, 2018, pp. 465–472, doi: 10.1007/978-3-319-99229-7_40.

[79] G. Jager, J. Schleiss, S. Usanavasin, S. Stober, and S. Zug, "Analyzing regions of safety for handling shared data in cooperative systems," in *Proc. 25th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Sep. 2020, pp. 628–635, doi: 10.1109/ETFA46521.2020.9211932.

[80] X. Lin, H. Zhu, R. Samanta, and S. Jagannathan, "Art: Abstraction refinement-guided training for provably correct neural networks," in *Proc. 20th Conf. Formal Methods Comput.-Aided Design, (FMCAD)*, Jan. 2020, pp. 148–157, doi: 10.34727/2020/isbn.978-3-85448-042-6_22.

[81] H. Zhao, X. Zeng, T. Chen, and Z. Liu, "Synthesizing barrier certificates using neural networks," in *Proc. 23rd Int. Conf. Hybrid Syst., Comput. Control*, Apr. 2020, pp. 1–11, doi: 10.1145/3365365.3382222.

[82] Q. Zhao, X. Chen, Y. Zhang, M. Sha, Z. Yang, W. Lin, E. Tang, Q. Chen, and X. Li, "Synthesizing ReLU neural networks with two hidden layers as barrier certificates for hybrid systems," in *Proc. 24th Int. Conf. Hybrid Syst., Comput. Control*, May 2021, pp. 1–11, doi: 10.1145/3447928.3456638.

[83] A. Peruffo, D. Ahmed, and A. Abate, "Automated and formal synthesis of neural barrier certificates for dynamical models," in *Proc. Int. Conf. Tools Algorithms Construct. Anal. Syst.*, Mar. 2021, pp. 370–388, doi: 10.1007/978-3-030-72016-2_20.

[84] M. Sha, X. Chen, Y. Ji, Q. Zhao, Z. Yang, W. Lin, E. Tang, Q. Chen, and X. Li, "Synthesizing barrier certificates of neural network controlled continuous systems via approximations," in *Proc. 58th ACM/IEEE Design Autom. Conf. (DAC)*, Dec. 2021, pp. 631–636, doi: 10.1109/DAC18074.2021.9586327.

[85] A. Claviere, E. Asselin, C. Garion, and C. Pagetti, "Safety verification of neural network controlled systems," in *Proc. 51st Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Jun. 2021, pp. 47–54, doi: 10.1109/DSN-W52860.2021.00019.

[86] Z. Wang, C. Huang, and Q. Zhu, "Efficient global robustness certification of neural networks via interleaving twin-network encoding," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2022, pp. 1087–1092.

[87] Q. Zhu, W. Li, H. Kim, Y. Xiang, K. Wardega, Z. Wang, Y. Wang, H. Liang, C. Huang, J. Fan, and H. Choi, "Know the unknowns: Addressing disturbances and uncertainties in autonomous systems," in *Proc. 39th Int. Conf. Comput.-Aided Design*, Nov. 2020, pp. 1–9, doi: 10.1145/3400302.3415768.

[88] R. Ivanov, J. Weimer, R. Alur, G. J. Pappas, and I. Lee, "Verisig: Verifying safety properties of hybrid systems with neural network controllers," in *Proc. 22nd ACM Int. Conf. Hybrid Syst., Comput. Control*, Apr. 2019, pp. 169–178, doi: 10.1145/3302504.3311806.

[89] R. Ivanov, T. Carpenter, J. Weimer, R. Alur, G. Pappas, and I. Lee, "Verisig 2.0: Verification of neural network controllers using Taylor model preconditioning," in *Proc. Int. Conf. Comput. Aided Verification*, vol. 12759, 2021, pp. 249–262, doi: 10.1007/978-3-030-81685-8_11.

[90] C. Sidrane, A. Maleki, A. Irfan, and M. J. Kochenderfer, "OVERT: An algorithm for safety verification of neural network control policies for nonlinear systems," *J. Mach. Learn. Res.*, vol. 23, no. 117, pp. 1–45, 2022. [Online]. Available: https://www.jmlr.org/papers/volume23/21-0847/21-0847.pdf

[91] H.-D. Tran, F. Cai, M. L. Diego, P. Musau, T. T. Johnson, and X. Koutsoukos, "Safety verification of cyber-physical systems with reinforcement learning control," *ACM Trans. Embedded Comput. Syst.*, vol. 18, no. 5s, pp. 1–22, Oct. 2019, doi: 10.1145/3358230.

[92] H. Fahmy, F. Pastore, and L. Briand, "HUDD: A tool to debug DNNs for safety analysis," in *Proc. IEEE/ACM 44th Int. Conf. Softw. Eng., Companion Proc. (ICSE-Companion)*, May 2022, pp. 100–104, doi: 10.1109/ICSE-Companion55297.2022.9793750.

[93] L. Pulina and A. Tacchella, "NeVer: A tool for artificial neural networks verification," *Ann. Math. Artif. Intell.*, vol. 62, nos. 3–4, pp. 403–425, Jul. 2011, doi: 10.1007/s10472-011-9243-0.

[94] G. Katz, "The marabou framework for verification and analysis of deep neural networks," in *Proc. Int. Conf. Comput. Aided Verification*, vol. 11561, 2019, pp. 443–452, doi: 10.1007/978-3-030-25540-4_26.

[95] C. Paterson, H. Wu, J. Grese, R. Calinescu, C. S. Păsăreanu, and C. Barrett, "DeepCert: Verification of contextually relevant robustness for neural network image classifiers," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 12852, 2021, pp. 3–17, doi: 10.1007/978-3-030-83903-1_5.

[96] C. E. Tuncali, J. Kapinski, H. Ito, and J. V. Deshmukh, "Reasoning about safety of learning-enabled components in autonomous cyber-physical systems," in *Proc. 55th Annu. Design Autom. Conf.*, Jun. 2018, pp. 1–6, doi: 10.1145/3195970.3199852.

[97] J. Val, R. Wisniewski, and C. S. Kallesoe, "Safe reinforcement learning control for water distribution networks," in *Proc. IEEE Conf. Control Technol. Appl. (CCTA)*, Aug. 2021, pp. 1148–1153, doi: 10.1109/CCTA48906.2021.9659138.

[98] D. T. Phan, R. Grosu, N. Jansen, N. Paoletti, S. A. Smolka, and S. D. Stoller, "Neural simplex architecture," in *Proc. 12th Int. Symp. NASA Formal Methods, (NFM)*, vol. 12229. Cham, Switzerland: Springer, 2020, pp. 97–114, doi: 10.1007/978-3-030-55754-6_6.

[99] D. Shukla, R. Lal, D. Hauptman, S. S. Keshmiri, P. Prabhakar, and N. Beckage, "Flight test validation of a safety-critical neural network based longitudinal controller for a fixed-wing UAS," in *Proc. AIAA AVIATION FORUM*, Jun. 2020, pp. 1–15, doi: 10.2514/6.2020-3093.

[100] S. Chen, Y. Sun, D. Li, Q. Wang, Q. Hao, and J. Sifakis, "Runtime safety assurance for learning-enabled control of autonomous driving vehicles," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 8978–8984, doi: 10.1109/ICRA46639.2022.9812177.

[101] Y. Peng, G. Tan, H. Si, and J. Li, "DRL-GAT-SA: Deep reinforcement learning for autonomous driving planning based on graph attention networks and simplex architecture," *J. Syst. Archit.*, vol. 126, May 2022, Art. no. 102505, doi: 10.1016/j.sysarc.2022.102505.

[102] J. Thumm and M. Althoff, "Provably safe deep reinforcement learning for robotic manipulation in human environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2022, pp. 6344–6350, doi: 10.1109/ICRA46639.2022.9811698.

[103] U. Mehmood, S. Sheikhi, S. Bak, S. A. Smolka, and S. D. Stoller, "The black-box simplex architecture for runtime assurance of autonomous CPS," in *Proc. 14th Int. Symp. NASA Formal Methods, (NFM)*, vol. 13260. Cham, Switzerland: Springer, 2022, pp. 231–250, doi: 10.1007/978-3-031-06773-0_12.

[104] M. Fazlyab, M. Morari, and G. J. Pappas, "Probabilistic verification and reachability analysis of neural networks via semidefinite programming," in *Proc. IEEE Conf. Decis. Control*, Dec. 2019, pp. 2726–2731, doi: 10.1109/CDC40024.2019.9029310.

[105] A. Grushin, J. Nanda, A. Tyagi, D. Miller, J. Gluck, N. C. Oza, and A. Maheshwari, "Decoding the black box: Extracting explainable decision boundary approximations from machine learning models for real time safety assurance of the national airspace," in *Proc. AIAA Scitech Forum*, Jan. 2019, p. 136, doi: 10.2514/6.2019-0136.

[106] R. Salay, M. Angus, and K. Czarnecki, "A safety analysis method for perceptual components in automated driving," in *Proc. IEEE 30th Int. Symp. Softw. Rel. Eng. (ISSRE)*, Oct. 2019, pp. 24–34, doi: 10.1109/ISSRE.2019.00013.

[107] R. Nahata, D. Omeiza, R. Howard, and L. Kunze, "Assessing and explaining collision risk in dynamic environments for autonomous driving safety," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 223–230, doi: 10.1109/ITSC48978.2021.9564966.

[108] S. Gerasimou, H. F. Eniser, A. Sen, and A. Cakan, "Importance-driven deep learning system testing," in *Proc. ACM/IEEE 42nd Int. Conf. Softw. Eng., Companion*, Oct. 2020, pp. 322–323, doi: 10.1145/3377812.3390793.

[109] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu, J. Zhao, and Y. Wang, "DeepGauge: Multi-granularity testing criteria for deep learning systems," in *Proc. 33rd ACM/IEEE Int. Conf. Automated Softw. Eng.*, Sep. 2018, pp. 120–131, doi: 10.1145/3238147.3238202.

[110] S. Wang, K. Pei, J. Whitehouse, J. Yang, and S. Jana, "Efficient formal safety analysis of neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 6367–6377. https://proceedings.neurips.cc/paper/2018/file/2ecd2bd94734e5dd392d8678bc64cdab-Paper.pdf

[111] K. Pei, Y. Cao, J. Yang, and S. Jana, "DeepXplore: Automated white-box testing of deep learning systems," *Commun. ACM*, vol. 62, no. 11, pp. 137–145, Oct. 2019, doi: 10.1145/3361566.

[112] Z. Chen, N. Narayanan, B. Fang, G. Li, K. Pattabiraman, and N. DeBardeleben, "TensorFI: A flexible fault injection framework for TensorFlow applications," in *Proc. IEEE 31st Int. Symp. Softw. Rel. Eng. (ISSRE)*, Oct. 2020, pp. 426–435, doi: 10.1109/ISSRE5003.2020.00047.

[113] Z. Chen, G. Li, K. Pattabiraman, and N. DeBardeleben, "BinFI: An efficient fault injector for safety-critical machine learning systems," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Nov. 2019, pp. 1–23, doi: 10.1145/3295500.3356177.

[114] Y. Jia, J. McDermid, T. Lawton, and I. Habli, "The role of explainability in assuring safety of machine learning in healthcare," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 4, pp. 1746–1760, Oct. 2022, doi: 10.1109/TETC.2022.3171314.

[115] S. Burton, "Safety assurance of machine learning for chassis control functions," in *Proc. Int. Conf. Comput. Saf., Rel. Secur.*, vol. 12852, 2021, pp. 149–162, doi: 10.1007/978-3-030-83903-1_10.

[116] G. Pedroza and A. Morayo, "Safe-by-design development method for artificial intelligent based systems," in *Proc. Int. Conf. Softw. Eng. Knowl. Eng.*, Jul. 2019, pp. 391–397, doi: 10.18293/SEKE2019-094.

[117] M. Douthwaite and T. Kelly, "Establishing verification and validation objectives for safety-critical Bayesian networks," in *Proc. IEEE Int. Symp. Softw. Rel. Eng. Workshops (ISSREW)*, Oct. 2017, pp. 302–309, doi: 10.1109/ISSREW.2017.60.

[118] I. Haring, F. Luttner, A. Frorath, M. Fehling-Kaschek, K. Ross, T. Schamm, S. Knoop, D. Schmidt, A. Schmidt, Y. Ji, Z. Yang, A. Rupalla, F. Hantschel, M. Frey, N. Wiechowski, C. Schyr, D. Grimm, M. R. Zofka, and A. Viehl, "Framework for safety assessment of autonomous driving functions up to SAE level 5 by self-learning iteratively improving control loops between development, safety and field life cycle phases," in *Proc. IEEE 17th Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Oct. 2021, pp. 33–40, doi: 10.1109/ICCP53602.2021.9733699.

[119] P. Koopman, U. Ferrell, F. Fratrik, and M. Wagner, "A safety standard approach for fully autonomous vehicles," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 11699, 2019, pp. 326–332, doi: 10.1007/978-3-030-26250-1_26.

[120] M. Mock, "An integrated approach to a safety argumentation for AI-based perception functions in automated driving," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 12853, 2021, pp. 265–271, doi: 10.1007/978-3-030-83906-2_21.

[121] 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, doi: 10.3390/make2040031.

[122] R. Salay and K. Czarnecki, "Improving ML safety with partial specifications," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 11699, 2019, pp. 288–300, doi: 10.1007/978-3-030-26250-1_23.

[123] A. Tarrisse and F. Massé, "Locks for the use of IEC 61508 to ML safety-critical applications and possible solutions," in *Proc. 31st Eur. Saf. Rel. Conf. (ESREL)*, 2021, pp. 3459–3466, doi: 10.3850/978-981-18-2016-8_661-cd.

[124] U. D. Ferrell and A. H. A. Anderegg, "Applicability of UL 4600 to unmanned aircraft systems (UAS) and urban air mobility (UAM)," in *Proc. AIAA/IEEE 39th Digit. Avionics Syst. Conf. (DASC)*, Oct. 2020, pp. 1–7, doi: 10.1109/DASC50938.2020.9256608.

[125] K. Aslansefat, I. Sorokos, D. Whiting, R. T. Kolagari, and Y. Papadopoulos, "SafeML: Safety monitoring of machine learning classifiers through statistical difference measures," in *Proc. Int. Symp. Model-Based Saf. Assessment*, vol. 12297, 2020, pp. 197–211, doi: 10.1007/978-3-030-58920-2_13.

[126] K. Aslansefat, W. Bridges, I. Sorokos, and D. Whiting. (Jun. 10, 2020). *GitHub—ISorokos/SafeML: Exploring Techniques for Estimating Safety of Machine Learning Classifiers*. Accessed: Nov. 24, 2022. [Online]. Available: https://github.com/ISorokos/SafeML

[127] M. Bergler, R. T. Kolagari, and K. Lundqvist, "Case study on the use of the SafeML approach in training autonomous driving vehicles," in *Proc. Int. Conf. Image Anal. Process.*, vol. 13233, 2022, pp. 87–97, doi: 10.1007/978-3-031-06433-3_8.

[128] T. Aoki, D. Kawakami, N. Chida, and T. Tomita, "Dataset fault tree analysis for systematic evaluation of machine learning systems," in *Proc. IEEE 25th Pacific Rim Int. Symp. Dependable Comput. (PRDC)*, Dec. 2020, pp. 100–109, doi: 10.1109/PRDC50213.2020.00021.

[129] J. F. Boulineau, "Safe recognition A.I. of a railway signal by onboard camera," in *Proc. 16th Eur. Dependable Comput. Conf. (EDCC)*, vol. 1279. Paris, France: Springer, 2020, pp. 5–19, doi: 10.1007/978-3-030-58462-7_1.

[130] L. Gauerhof, R. Hawkins, C. Picardi, C. Paterson, Y. Hagiwara, and I. Habli, "Assuring the safety of machine learning for pedestrian detection at crossings," in *Proc. 39th Int. Conf. Comput. Saf., Rel. Secur., (SAFECOMP)*, vol. 12234. Renningen, Germany: Springer, 2020, pp. 197–212, doi: 10.1007/978-3-030-54549-9_13.

[131] M. Klaes, R. Adler, I. Sorokos, L. Joeckel, and J. Reich, "Handling uncertainties of data-driven models in compliance with safety constraints for autonomous behaviour," in *Proc. 17th Eur. Dependable Comput. Conf. (EDCC)*, Sep. 2021, pp. 95–102, doi: 10.1109/EDCC53658.2021.00021.

[132] A. Subbaswamy, R. Adams, and S. Saria, "Evaluating model robustness and stability to dataset shift," in *Proc. 24th Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 130, 2021, pp. 2611–2619.

[133] J. Firestone and M. B. Cohen, "The assurance recipe: Facilitating assurance patterns," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 11094, 2018, pp. 22–30, doi: 10.1007/978-3-319-99229-7_3.

[134] J. Bragg and I. Habli, "What is acceptably safe for reinforcement learning?" in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 11094, 2018, pp. 418–430, doi: 10.1007/978-3-319-99229-7_35.

[135] L. Gauerhof, P. Munk, and S. Burton, "Structuring validation targets of a machine learning function applied to automated driving," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 11093, 2018, pp. 45–58, doi: 10.1007/978-3-319-99130-6_4.

[136] C.-H. Cheng, C.-H. Huang, and G. Nuhrenberg, "Nn-dependability-kit: Engineering neural networks for safety-critical autonomous driving systems," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2019, pp. 1–6, doi: 10.1109/ICCAD45719.2019.8942153.

[137] C.-H. Cheng and R. Yan, "Continuous safety verification of neural networks," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Feb. 2021, pp. 1478–1483, doi: 10.23919/DATE51398.2021.9473994.

[138] *Railway Applications Communication, Signalling and Processing Systems Software for Railway Control and Protection Systems*, document CENELEC, EN50128:2011, Brussels, Belgium, 2011.

[139] S. Gupta, I. Ullah, and M. G. Madden, "Coyote: A dataset of challenging scenarios in visual perception for autonomous vehicles," in *Proc. CEUR Workshop*, vol. 2916, 2021, pp. 1–9.

[140] C. J. Hong and V. R. Aparow, "System configuration of Human-in-the-loop simulation for level 3 autonomous vehicle using IPG CarMaker," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2021, pp. 215–221, doi: 10.1109/IoTaIS53735.2021.9628587.

[141] A. Mjeda and G. Botterweck, "Uncertainty entangled; modelling safety assurance cases for autonomous systems," *Electron. Commun. EASST*, vol. 79, pp. 1–10, May 2020. [Online]. Available: https://journal.ub.tu-berlin.de/eceasst/article/download/1124/1072

[142] A. Corso and M. J. Kochenderfer, "Transfer learning for efficient iterative safety validation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 8, May 2021, pp. 7125–7132. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16876/16683

[143] C. Smith, E. Denney, and G. Pai, "Hazard contribution modes of machine learning components," in *Proc. CEUR Workshop*, vol. 2560, 2020, pp. 14–22.

[144] H. Barzamini, M. Shahzad, H. Alhoori, and M. Rahimi, "A multi-level semantic web for hard-to-specify domain concept, pedestrian, in ML-based software," *Requirements Eng.*, vol. 27, no. 2, pp. 161–182, Jun. 2022, doi: 10.1007/s00766-021-00366-0.

[145] J. H. Husen, H. Washizaki, H. T. Tun, N. Yoshioka, Y. Fukazawa, and H. Takeuchi, "Traceable business-to-safety analysis framework for safety-critical machine learning systems," in *Proc. 1st Int. Conf. AI Eng., Softw. Eng. (AI)*, May 2022, pp. 50–51, doi: 10.1145/3522664.3528619.

[146] R. Alexander and T. Kelly, "Supporting systems of systems hazard analysis using multi-agent simulation," *Saf. Sci.*, vol. 51, no. 1, pp. 302–318, Jan. 2013, doi: 10.1016/j.ssci.2012.07.006.

[147] W. Ruan, X. Huang, and M. Kwiatkowska, "Reachability analysis of deep neural networks with provable guarantees," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2651–2659.

[148] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," *Artif. Intell.*, vol. 279, Feb. 2020, Art. no. 103201, doi: 10.1016/j.artint.2019.103201.

[149] H. Lin and W. Liu, "Risks and prevention in the application of AI," in *Proc. Int. Conf. Mach. Learn. Big Data Anal. IoT Secur. Privacy*, vol. 1283, 2021, pp. 700–704, doi: 10.1007/978-3-030-62746-1_104.

[150] P. Sarathy, S. Baruah, S. Cook, and M. Wolf, "Realizing the promise of artificial intelligence for unmanned aircraft systems through behavior bounded assurance," in *Proc. IEEE/AIAA 38th Digit. Avionics Syst. Conf. (DASC)*, Sep. 2019, pp. 1–8, doi: 10.1109/DASC43569.2019.9081649.

[151] A. Causevic, A. V. Papadopoulos, and M. Sirjani, "Towards a framework for safe and secure adaptive collaborative systems," in *Proc. IEEE 43rd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Jul. 2019, pp. 165–170, doi: 10.1109/COMPSAC.2019.10201.

[152] S. Seon and J.-W. Kim, "Designing a modular safety certification system for convergence products–focusing on autonomous driving cars," *J. Korean Soc. Quality Manag.*, vol. 46, no. 4, pp. 1001–1014, 2018. [Online]. Available: https://koreascience.kr/article/JAKO201816842430631.page

[153] S. Anastasi, M. Madonna, and L. Monica, "Implications of embedded artificial intelligence–machine learning on safety of machinery," *Proc. Comput. Sci.*, vol. 180, pp. 338–343, Jan. 2021, doi: 10.1016/j.procs.2021.01.171.

[154] B. B. Gupta, A. Gaurav, E. C. Marin, and W. Alhalabi, "Novel graph-based machine learning technique to secure smart vehicles in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, early access, May 30, 2022, doi: 10.1109/TITS.2022.3174333.

[155] Linux Foundation AI & Data Foundation. (Nov. 15, 2022). *GitHub Trusted-AI/Adversarial-Robustness-Toolbox: Adversarial Robustness Toolbox (ART) Python Library for Machine Learning Security Evasion, Poisoning, Extraction, Inference Red and Blue Teams*. Accessed: Nov. 23, 2022. [Online]. Available: https://github.com/Trusted-AI/adversarial-robustness-toolbox

[156] S. Alemany, J. Nucciarone, and N. Pissinou, "Jespipe: A plugin-based, open MPI framework for adversarial machine learning analysis," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 3663–3670, doi: 10.1109/BIGDATA52589.2021.9671385.

[157] F. R. Ward and I. Habli, "An assurance case pattern for the interpretability of machine learning in safety-critical systems," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, 2020, pp. 395–407, doi: 10.1007/978-3-030-55583-2_30.

[158] I. Ramezani, K. Moshkbar-Bakhshayesh, N. Vosoughi, and M. B. Ghofrani, "Applications of soft computing in nuclear power plants: A review," *Prog. Nucl. Energy*, vol. 149, Jul. 2022, Art. no. 104253, doi: 10.1016/j.pnucene.2022.104253.
[159] S. Iqbal, T. M. Khan, K. Naveed, S. S. Naqvi, and S. J. Nawaz, "Recent trends and advances in fundus image analysis: A review," *Comput. Biol. Med.*, vol. 151, Dec. 2022, Art. no. 106277, doi: 10.1016/J.COMPBIOMED.2022.106277.
[160] T. M. Khan and A. Robles-Kelly, "Machine learning: Quantum vs classical," *IEEE Access*, vol. 8, pp. 219275–219294, 2020, doi: 10.1109/ACCESS.2020.3041719.
[161] Y. Kwak, W. J. Yun, S. Jung, J.-K. Kim, and J. Kim, "Introduction to quantum reinforcement learning: Theory and PennyLane-based implementation," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2021, pp. 416–420, doi: 10.1109/ICTC52510.2021.9620885.
[162] Scikit-Learn. (2022). *Scikit-Learn: Machine Learning in Python Scikit-Learn 1.1.2 Documentation*. Accessed: Oct. 7, 2022. [Online]. Available: https://scikit-learn.org/stable/
[163] E. Wozniak, C. Carlan, E. Acar-Celik, and H. J. Putzer, "A safety case pattern for systems with machine learning components," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.*, vol. 12235, 2020, pp. 370–382, doi: 10.1007/978-3-030-55583-2_28.
[164] A. V. da Silva Neto, L. F. Vismari, R. A. V. Gimenes, D. B. Sesso, J. R. de Almeida, P. S. Cugnasca, and J. B. Camargo, "A practical analytical approach to increase confidence in PLD-based systems safety analysis," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3473–3484, Dec. 2018, doi: 10.1109/JSYST.2017.2726178.
[165] A. V. S. Neto and P. S. Cugnasca, "Technical research report—Remarks on the systematic literature review on the safety assurance of AI-based systems—Version 6," Saf. Anal. Group (GAS), Dept. Comput. Eng. Digit. Syst., Escola Politécnica, Universidade de São Paulo (USP), São Paulo, Brazil, 2022, doi: 10.5281/zenodo.7358711.

ANTONIO V. SILVA NETO was born in São Paulo, Brazil, in 1988. He received the bachelor's degree in electrical engineering and the M.Sc. degree from the School of Engineering, University of São Paulo (Poli-USP), in 2010 and 2014, respectively, where he is currently pursuing the D.Sc. degree with the Safety Analysis Group (GAS), working on methods for the safety assurance of artificial intelligence-based systems, supported by the Brazilian institutions CAPES (Coordenação de it Aperfeiçoamento de Pessoal de Nível Superior) and FDTE (Fundação para o Desenvolvimento Tecnológico da Engenharia).

During his D.Sc. degree, he has also been a Teaching Assistant with the Digital Laboratory undergraduate courses offered as part of the computer engineering undergraduate courses, since 2021. Before starting his D.Sc. degree, he was with Alstom Brazil, in 2013 and (2018–2020), respectively, where he acted as the Safety Assurance Manager for research & development projects and a Safety Assurance Engineer for metro signaling projects on Brazil, Mexico, Panama, and Chile. Moreover, he was also with FDTE (2009–2018), where he was an Independent System Safety Analyst for Brazilian safety-critical projects on metro and air traffic control domains. He has five papers in scientific journals, six conference papers, and 11 participations on examination boards of Poli-USP electrical and computer engineering undergraduate dissertations.

Mr. Silva Neto is also a Reviewer for the IEEE Systems Journal.

JOÃO B. CAMARGO JR. received the bachelor's degree in electronic engineering and the M.Sc. and Ph.D. degrees from the School of Engineering, University of São Paulo (Poli-USP), São Paulo, Brazil, in 1981, 1989, and 1996, respectively.

He is currently an Associate Professor with the Department of Computer Engineering and Digital Systems (PCS), Poli-USP, where he is also the Coordinator of the Safety Analysis Group (GAS). He has 40 articles in scientific journals, five organized books, six published book chapters, and 115 complete works published in proceedings of conferences. He is a Reviewer in different scientific journals, such as the IEEE Transactions on Intelligent Transportation Systems, *Journal of Intelligent and Robotic Systems*, *Risk Analysis*, the *Journal of Advanced Transportation*, *Reliability Engineering and System Safety*, *Transportation Research—Part C*, and the IEEE Software.

JORGE R. ALMEIDA JR. received the bachelor's degree in electronic engineering and the M.Sc. and Ph.D. degrees from the School of Engineering, University of São Paulo (Poli-USP), São Paulo, Brazil, in 1981, 1989, and 1995, respectively.

He is currently an Associate Professor with the Department of Computer Engineering and Digital Systems (PCS), Poli-USP, where he is also a member of the Safety Analysis Group (GAS). His research interests include reliable and safe computational systems for critical application.

PAULO S. CUGNASCA received the bachelor's degree in electronic engineering and the M.Sc. and Ph.D. degrees from the School of Engineering, University of São Paulo (Poli-USP), São Paulo, Brazil, in 1987, 1993, and 1999, respectively.

He is currently an Associate Professor with the Department of Computer Engineering and Digital Systems (PCS), Poli-USP, where he is also a member of the Safety Analysis Group (GAS). His research interests include reliable and safe computational systems for critical application.