frontiers | Frontiers in Artificial Intelligence

TYPE Systematic Review
PUBLISHED 03 October 2022
DOI 10.3389/frai.2022.952773

OPEN ACCESS

EDITED BY
Hamed Zolbanin,
University of Dayton, United States

REVIEWED BY
Mostafa Haghi Kashani,
Islamic Azad University, Iran
Rabia Saleem,
University of Derby, United Kingdom
Mohammad Mahdi Amirian,
Amirkabir University of Technology,
Iran

*CORRESPONDENCE
Moritz Wäschle
moritz.waeschle@kit.edu

SPECIALTY SECTION
This article was submitted to
AI in Business,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 25 May 2022
ACCEPTED 08 September 2022
PUBLISHED 03 October 2022

CITATION
Wäschle M, Thaler F, Berres A,
Pölzlbauer F and Albers A (2022) A
review on AI Safety in highly
automated driving.
Front. Artif. Intell. 5:952773.
doi: 10.3389/frai.2022.952773

A review on AI Safety in highly automated driving

Moritz Wäschle[1]*, Florian Thaler[2], Axel Berres[3],
Florian Pölzlbauer[2] and Albert Albers[1]

[1]IPEK—Institute of Product Engineering, ASE—Advanced Systems Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [2]Virtual Vehicle Research GmbH, Graz, Austria, [3]German Aerospace Center, Cologne, Germany

Remarkable progress in the fields of machine learning (ML) and artificial intelligence (AI) has led to an increased number of applications of (data-driven) AI systems for the partial or complete control of safety-critical systems. Recently, ML solutions have been particularly popular. Such approaches are often met with concerns regarding their correct and safe execution, which is often caused by missing knowledge or intransparency of their exact functionality. The investigation and derivation of methods for the safety assessment of AI systems are thus of great importance. Among others, these issues are addressed in the field of AI Safety. The aim of this work is to provide an overview of this field by means of a systematic literature review with special focus on the area of highly automated driving, as well as to present a selection of approaches and methods for the safety assessment of AI systems. Particularly, validation, verification, and testing are considered in light of this context. In the review process, two distinguished classes of approaches have been identified: On the one hand established methods, either referring to already published standards or well-established concepts from multiple research areas outside ML and AI. On the other hand newly developed approaches, including methods tailored to the scope of ML and AI which gained importance only in recent years.

KEYWORDS

AI Safety, systematic literature review, highly automated driving, value alignment, adversarial robustness

1. Introduction

Background and motivation

In the field of highly automated driving, non-linear system behavior and an unknown environment can be addressed by AI systems. These systems need to act safely at all times – therefore, AI Safety is an important research need. For many AI systems, especially for systems based on neural networks, classical safety analysis methods can hardly be applied. Moreover, current standards do not address the development of safe AI systems. Consequently, for the safety assessment of AI systems the activities validation, verification, and testing need to be considered. In general, safety is particularly of interest dealing with technologies. The underlying philosophical question: "What risk are we willing to take when we use technology?" has led to the establishment of corresponding targets and process specifications in the safety standards (MIT, 2021). The question of trust in the safety of a system is susceptible if it is controlled—fully or partially—by an AI controller. In recent years, in particular, data-driven AI solutions gained importance for controlling safety-critical systems. This also applies to the area of automated driving, where driver-less vehicles operated by machine learning techniques are already a reality, as shown in Krafcik (2020).

Citing Rudner and Toner (2021), machine learning systems do not come with safety guarantees and there are no means yet to ensure that these systems can be operated with a very small risk of failure. Hence, methods providing safety of AI systems and approaches for their validation and verification are needed. The following questions arise: How can AI applications adequately become safer and resilient against undetected development errors and system failures during operation? Can conventional methods be used to develop safe AI applications—in which use cases do new solutions need to be found? Regarding the latter, there are use cases where conventional algorithms can be used in areas wherein an AI solution can be applied safely. As an example, consider a friction clutch. The pressure allowed to close it and whether it can be closed at all can be determined by conventional algorithms. An AI approach could then be designed to find an optimal filling process in terms of comfort and wear of the gearshift. In many use cases, such a dual approach can hardly be applied. For example, if we think of trajectory planning, it seems difficult to monitor an AI solution with a conventional approach. The number of different ways to plan trajectories is too large, and the problem itself too general. AI solutions and classical solutions are combinable and can be mutual beneficial, but it is challenging to identify suitable applications allowing both approaches to fulfill their potential.

Contribution

The contribution of this paper is an overview of the current research of AI Safety. In order to identify the most relevant categories and topics of this field, a systematic literature review following (Brereton et al., 2007; Kitchenham and Charters, 2007; Liberati et al., 2009; Schumann et al., 2020) was applied. Based on the results of this review process, two categories are identified and discussed: the so-called classical approaches and the new approaches. The section on classical approaches examines ways of how established norms and standards address the problem class of AI Safety. Under new approaches concerning validation, verification, and testing, we summarize newly developed approaches for the safety assessment tailored to the needs of AI use cases. In addition, models for the robust development of AI systems are taken into account.

Outline

This paper is divided into nine sections (see Figure 1). After the "Introduction" in Section 1, Section 2 examines the main research areas "Highly automated driving" and "AI safety" and links to already existing contributions in these fields. Section 3 considers already published literature reviews for the defined problem class of AI safety of highly automated driving. Section 4 describes the planning and conducting of the "systematic literature review," which forms the basis of this review paper. Sections 5, 6 deal with "classical and new approaches" in the field of AI Safety based on the elaborated research questions. In Sections 7, 8, the findings of the provided literature review are discussed, and topics that, in the opinion of the authors, require further research are mentioned. Finally, Section 9 summarizes the contribution and limitations of this paper.

2. Background

2.1. Highly automated driving

Over the last years, there have still been many traffic accidents. Alone in the United States of America, almost 40,000 lives were lost in 2020 (National Center for Statistics and Analysis, 2017; Stewart, 2022). The automation of driving tasks can improve traffic safety while improving traffic efficiency, environmental impact, and comfort. However, classical safety approaches do not consider Artificial Intelligence in systems. According to the system theory described by Ropohl (2009), a system acts in a context in order to fulfill its purpose. In Systems Engineering, the purpose and understanding of the context are important tasks (Walden et al., 2015). Then, a system solution is developed that can fulfill the purpose. Models of the complex automotive system support its development. However, the models cannot be complete due to the complex dependencies and the linearization of the behavior. Here, two potential classes of problems emerge that can be more effectively solved by AI. These are dealing with non-linear behavior and an unknown environment.

Problem 1: Non-linear behavior: This problem class can describe all complicated and complex problems. Schroeder (2010) describes the possible application of neural networks for non-linear systems. However, systems having non-linear behavior are classical problems. Classical problems consider a known system, e.g., a clutch controlling the maximum contact pressure. In these systems, both the purpose and the context are well-defined and do not change. Furthermore, for these problems, the systems have been continuously optimized over the years.

Due to existing non-linear behavior resulting from the underlying physics or the interaction of different system parts, the question arises of whether AI algorithms can improve control.

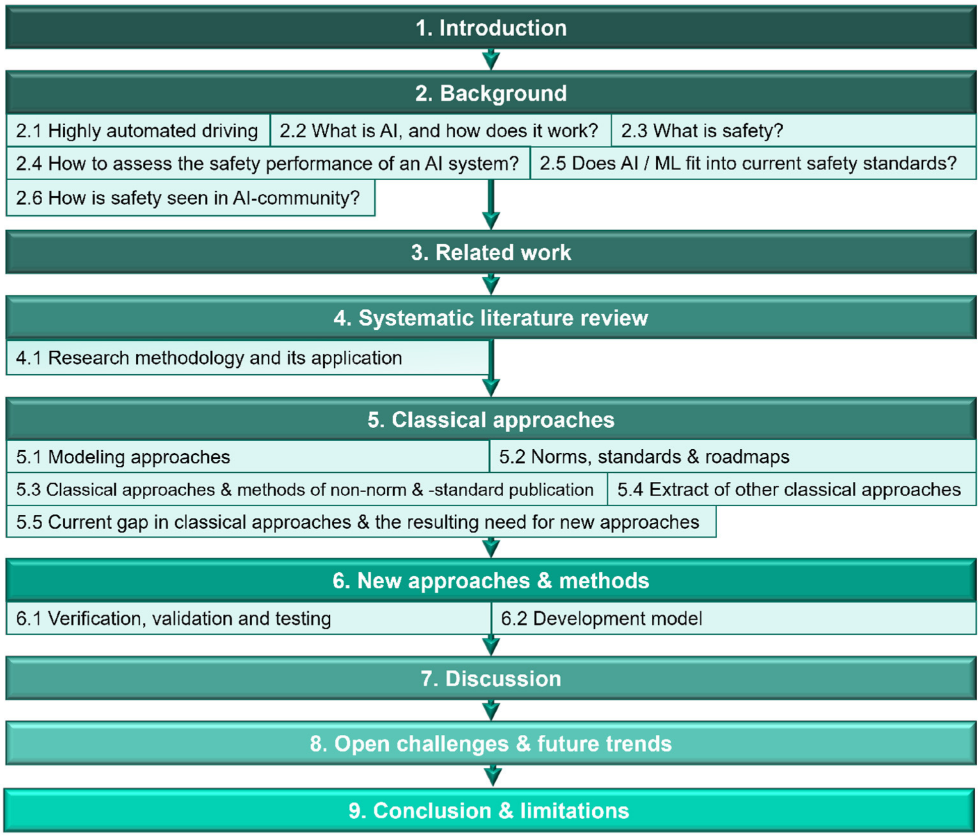
1. Introduction
2. Background
2.1 Highly automated driving
2.2 What is AI, and how does it work?
2.3 What is safety?
2.4 How to assess the safety performance of an AI system?
2.5 Does AI / ML fit into current safety standards?
2.6 How is safety seen in AI-community?
3. Related work
4. Systematic literature review
4.1 Research methodology and its application
5. Classical approaches
5.1 Modeling approaches
5.2 Norms, standards & roadmaps
5.3 Classical approaches & methods of non-norm & -standard publication
5.4 Extract of other classical approaches
5.5 Current gap in classical approaches & the resulting need for new approaches
6. New approaches & methods
6.1 Verification, validation and testing
6.2 Development model
7. Discussion
8. Open challenges & future trends
9. Conclusion & limitations

FIGURE 1
Structure of the review.

Problem 2: Unknown Environment: This class deviates from the classical approach. In this problem class, the system's purpose is clearly defined, but the system's environment may change.

One example is an in-car lane departure assistant. The task of this system is to keep the car in the specified lane automatically. In doing so, the system must be able to deal with different weather conditions and road markings. Since here, the testing of the function can only be done partially due to the pervasive environmental conditions, not all failures can be identified. Thus, this problem class describes the operation of an unsafe system in an unsafe environment.

To conclude, the problem classes of non-linear behavior and unknown environment illustrate the need for AI systems. These systems need to act safely; therefore, *AI Safety* is an important research need.

2.2. What is AI, and how does it work?

According to Nilsson (2009) and Russell and Norvig (2022), Artificial Intelligence (AI) is the subfield of computer science dealing with the task of building intelligent entities, which means to enable these entities to function appropriately and with foresight in their environment.

In recent years, data-driven AI solutions, or machine learning (ML) solutions[1], have gained particular importance. Machine learning deals with systems that use mathematical and statistical methods to extract regularities and patterns from large amounts of data to solve complex decision or control problems. The process of deriving or as well improving a decision or classification rule is called training. Mostly, training corresponds to solving a mathematical optimization problem for

[1] In the course of this work, the terms machine learning solutions and data-driven solutions are used synonymously. Since machine learning is based on the extraction of regularities and patterns from data-sets, every machine learning solution can also be classified as data-driven. Conversely, there are data-driven methods, for example, in the field of statistics, which cannot be associated with machine learning. Consequently, the terms are to be distinguished in general. In the given context, however, in the opinion of the authors, there is no confusion regarding the meaning of the terms, and therefore the synonymous use of these terms is permissible.

data adaptation by (numerically) minimizing or maximizing a loss function.

According to Sutton and Barto (2018) and Frochte (2019), the field of ML can be subdivided into three broad categories: supervised learning, unsupervised learning, and reinforcement learning[2]. In the supervised learning paradigm, ground truth in terms of labeled data is used for the training of the learning system. The objective is to learn the relationships between the data and its labels in such a way, that the system is able to label accurately unseen data. In the field of autonomous driving, solutions based on this approach are, for example, used for tasks like traffic sign detection, pedestrian detection, or road marking detection (Bachute and Subhedar, 2021). Unlike supervised learning, unsupervised learning does not rely on labeled data. Methods belonging to this field seek for hidden and previously undetected patterns or groupings inside a given data set without prior information on the data. In the automotive sector this type of machine learning can be used for example, as shown in Li et al. (2018), for the clustering of vehicle encounter data. The training principle of reinforcement learning techniques is based on the principle of learning through interaction. Due to repeated interaction with its environment, the learning system discovers which actions generate positive or negative feedback to solve a given task. The system is then encouraged to derive a strategy that generates maximal feedback in terms of a numerical reward function. As suggested by Kober et al. (2013), reinforcement learning approaches are particularly suitable for application in the field of robotics. Furthermore, as shown in Folkers et al. (2019), these methods can be used to derive a controller for a self-driving car.

For the mathematical representation of machine learning systems, so-called artificial neural networks (ANN) are often used. From a purely mathematical viewpoint, any ANN is a parameter-dependent function approximator allowing to approximate a broad class of functions arbitrarily accurate (Hornik, 1991). Especially this feature is fundamental for their broad and frequent use. In terms of their design, structure, and functioning ANNs mimic human brains, which also explains their naming. For a detailed description of the concept, see Hecht-Nielsen (1992) and Abdi (1994).

2.3. What is safety?

Colloquially, safety is understood as a state of freedom from risk or danger. In technology, several definitions of the term safety can be found:

1. Following ISO Central Secretary, International Electrotechnical Commission (2014), the term safe describes a state to protect against recognized hazards which likely cause harm.
2. In ISO 61508 (International Electrotechnical Commission, 2010), safety is defined as freedom from unacceptable risks. An acceptable or tolerable risk refers to a risk that is tolerated in a predefined context on the basis of current society values.
3. According to MIL 882E (Department of Defense Systems (DoD), 2012), safety is understood as freedom from states which might cause injury, death, illness, damage to or loss of property or equipment or environment.
4. The standard ISO 26262—see ISO Central Secretary (2018c)—focuses on functional safety in the automotive context. It describes functional safety with the lack of unreasonable risk caused by hazards that are caused by the malfunctions of E/E systems.

According to these definitions, one may say that a system is safe if it can be operated free from all the identified and non-tolerable hazards. This definition also applies to systems comprising AI algorithms, making it necessary to assess their safety. Since many AI systems are built upon ANNs and established safety analysis procedures are hardly applicable in this case—see Section 6 for a more detailed discussion - these methods have to be adapted, or novel approaches have to be found.

2.4. How to assess the safety performance of an AI system?

Proving safety for AI systems in a rigorous way is a difficult task and still an open problem—see, for instance, Section 6. This makes verification and validation approaches to ensuring safety requirements all the more important.

2.4.1. Verification

As pointed out in Gausemeier and Moehringer (2002), verification colloquially is the answer to the question: Is the correct product being developed? Technically, according to ISO Central Secretary (2015), verification deals with the task of confirming through the provision of objective evidence that requirements have been satisfied. Thus, following Fisher (2007) and Goodfellow and Papernot (2017), verification aims to give confidence that the product was built adequately and that it will not misbehave under a vast range of circumstances.

[2] In the context of this paper we only aim to give a rough overview of the field of ML and its subdivisions. Therefore, we content ourselves with dividing the field into these three categories. For the sake of completeness, we would like to point out that a finer subdivision can be made. In Russell and Norvig (2022), and also in Sarker (2021) next to supervised learning, unsupervised learning, and reinforcement learning, the additional sub-field of semi-supervised learning is mentioned—in Ayodele (2010) an even finer subdivision is introduced. For an insight into these additional sub-areas, we refer to the literature cited.

Usually, the process of verification is realized formally. However, for the verification of ML systems, alternative methods are required (see Section 6).

2.4.2. Validation

Informally, by Gausemeier and Moehringer (2002), validation seeks an answer to the question: Is the right product being developed? In a technical context, validation deals with the issue of confirming by providing objective evidence that the requirements for an intended application have been satisfied. In contrast to verification, systems are not validated formally. Rather, appropriate tests are designed and executed for the process of validation.

2.4.3. Testing

Following Ebel (2015), by testing we understand the determination of properties of a system. In particular, testing provides information about the system, which can be used to check whether the system satisfies defined requirements, objectives, or hypotheses entirely, partially, or not.

2.5. Does AI/ML fit into current safety standards?

One might be tempted to view machine learning as just a novel paradigm for designing and implementing software components for cyber-physical systems. So the question arises:

Can ML-based systems be designed, verified, and certified according to current safety standards?

Salay et al. (2017) analyzed to which degree ML-based systems could satisfy ISO26262. They conclude that while a large portion of the standard could be satisfied, there exists a set of open issues:

1. ML can create new hazards not due to malfunctioning of the component but due to the complex interaction with humans.
2. Due to its novel development cycle, ML-based systems have distinct faults and failure modes.
3. The capabilities of an ML-based system are inherently tied to the quality of the training data-set. However, this data set is—by definition—incomplete.
4. ML systems having a black box character, e.g., systems based on ANNs, violate the call for hierarchical decomposition.
5. ISO26262 mandates specific techniques for software design, verification, and validation. Some of which are only valid for imperative programming languages.

Consequently, we have to conclude that **ML-based systems cannot fully satisfy current safety standards**. In 2018 ISO initiated a standardization project toward AI: ISO/IEC JTC 1/SC 42 Artificial intelligence. Within this project, working group 3 (WG3) focuses on *trustworthiness*. One aspect of the WG3 is to investigate approaches to realize AI systems' *safety* as well as robustness, reliability, resiliency, accuracy, and privacy. In addition, WG3 has a project on AI risk management, which aims at a standard to address certification processes. A good overview of the current state of "AI standardization" within ISO/IEC JTC 1/SC 42 can be found in Zielke (2020). ISO/IEC AWI TR 5469 (Artificial intelligence—Functional Safety and AI systems) addresses precisely the issue of safety. Also, ISO26262 is currently working on its 3rd edition, in which AI/ML will be addressed. Besides these efforts toward standardization, for several essential topics (especially concerning "safety"), the scientific basis is not sufficiently solid yet, as concluded in Zielke (2020) with three current issues:

- "Formal methods for the verification of neural networks or for the assessment of their robustness (Zielke, 2020)" (c.f. Huang et al., 2020).
- "Architectures and training methods for robust solutions based on deep neural networks (Zielke, 2020)" (c.f. Becker et al., 2020).
- "Methods and tools for generating comprehensible explanations for AI-based decision processes (Zielke, 2020)" (c.f. Goebel et al., 2018).

One key aspect of developing safety-critical systems is the assurance case. The safety argument proves that the system is safe and essential for certification. Schwalbe and Schels (2020) highlight a set of key challenges to overcome in order to assure the safety of ML-based systems: (1) powerful solvers, (2) use of expert knowledge, (3) validation of data and model diversity, (4) model introspection with guarantees. The authors highlight the challenges along the safety life cycle and provide a detailed table listing promising approaches and open challenges.

UL4600 (Underwriters Laboratories, 2020) addresses the safety of autonomous driving systems without human intervention. Therefore, the standard relies on a claim-based approach by using assurance cases. Koopman et al. (2019a) build on the UL4600 standard and derives an approach with goal-based safety cases, and feedback loops in the context of autonomous driving.

2.6. How is safety seen in AI-community?

The AI community uses the term *AI Safety*. As summarized in Berlinic (2019), AI Safety may conclude that AI is beneficial or detrimental. AI Safety needs research work to ensure it is beneficial. Yampolskiy and Fox (2012) define the term Safety Engineering for Artificial General Intelligence with the aim of creating safe systems. Amodei et al. (2016) describe AI Safety by mitigating accident risk in machine learning systems. Russell et al. (2015) explain safety by complying with the terms verification, validity[3], security, and control. Besides safety the terms trustworthiness and confidence need to be considered. According to the mission of the Stanford Center for AI Safety, systems with AI must be safe and trustworthy to facilitate their application in society (Barrett et al., 2022).

In the understanding of this paper, AI Safety deals with the interaction in operating systems to ensure a safe operation (cf. Yampolskiy and Fox, 2012; Amodei et al., 2016. Safety mechanisms help to come to safe operating states. Trustworthiness, verification, validation, security, and control need to be considered in the field of AI Safety (cf. Russell et al., 2015; Barrett et al., 2022).

In the following, seven main challenges in the context of beneficial AI are described (Berlinic, 2019):

- **Fairness**: Machine learning uses decision-making, which might be biased. For instance, the data set is biased due to human prejudices. "AI safety asks: *How do we build AI that is unbiased and does not systematically discriminate against underprivileged groups?" (Berlinic, 2019).*
- **Transparency**: In many cases it is difficult to understand how ML systems make decisions. Especially when the ML system includes a neural network, there is a lack of explainability of the decisions made by the system. "AI safety asks: *How do we build AI that can explain its decisions? How do we build AI that can explain why it made the wrong decision?" (Berlinic, 2019)* (c.f. research field Explainable AI).
- **Misuse**: The algorithms can be maliciously used by people. "AI safety asks: *How do we ensure that AI is only used for good causes?" (Berlinic, 2019).*
- **Security**: AI, like every software, is vulnerable to malicious attacks. This might result in unintended actions of the initial design purpose. "AI safety asks: *How do we prevent malicious actors from abusing imperfect AI Systems?" (Berlinic, 2019).*
- **Policy**: AI has an increasing impact on products and society. "AI safety asks: *How do we ensure that AI benefits all, not only a few? How do we handle the disruptions that will be caused by its development?" (Berlinic, 2019).*
- **Ethics**: AI needs to act under certain ethical standards. Human values are one of the broader goals to limit functionalities. "AI safety asks: *How do we decide the values that AI promotes?" (Berlinic, 2019).*
- **Control/Alignment**: AI must be aligned with the values of the designer so that no misinterpretation can happen. "AI safety asks: *How do we align AI with our values so that it does what we intend, not what we ask?" (Berlinic, 2019)* (c.f. research field Value Alignment).

TABLE 1 Overview of related publications.

Type	Publication	Main topic	Covered years
Survey	Juric et al., 2020	Quantitative review with future trends	1985–2019
Report	European Commission, 2020	Implications of legislations concerning AI	Not defined
Blog	Dawson, 2017	Application to critical infrastructures	Not defined
Blog	Krakovna, 2021	AI alignment	Not defined

As can be seen, AI Safety is a broad topic that exceeds the scope of a single paper. Hence, in this publication, we will focus on the engineering question: How can AI-based systems be designed and executed so that they do not cause accidents? Here, the term accident is defined as unintended and harmful behavior that may emerge from poor AI design (Amodei et al., 2016). According to Amodei et al. (2016) accidents caused by AI-based systems mainly stem from three issues:

- Having the wrong objective function. The wrong function can result from typical mechanisms like negative side effects or reward hacking (i.e., algorithms quickly use bucks to get unintended rewards).
- Having an objective function which is not affordable to evaluate it frequently. For instance, a cleaning robot might not know how to handle each possible tiny object (suck it in or leave it).
- Undesirable behavior throughout the learning process, e.g., from insufficient training data. For instance, the vehicle's environment is changing with new devices and infrastructure to handle.

3. Related work

Table 1 gives an overview of the related works considered in this publication.

In Juric et al. (2020), the authors review the topic of "AI Safety" in a quantitative manner. *Via* a dedicated list of keywords, they queried literature databases (SCOPUS, Web of Science, Google Scholar) in order to see how these topics evolved from 1985 to 2019. The single largest growth is in *interpretability*. Strong growth in the number of publications is also in *AI ethics* and *adversarial robustness*, medium growth is in *value alignment* and *safe exploration*, while only slight growth is in *fairness* and *privacy*. Emerging topics seem to be *distribution shift*, *safe exploration*, *interruptibility*, and *reward hacking*. However, a substantial number of ideas in the area of AI Safety is not peer-reviewed

[3] The authors use the term validity, however in this work we will use the term validation for describing that a system fulfills its purpose.

published. Nonetheless, a set of future research directions can be seen:

- "One of the most important open problems in **explainability** is that there is no agreement on what an explanation is Juric et al. (2020)." The evaluation of the understanding of explanations to humans is not well explored, and there is still no algorithm that provides both high accuracy and explainability.
- Learning a reward function is essential for the future of the field of **value alignment**. In addition, value discovery is another major topic in this research direction. It is the identification of better reward functions with aligned algorithms.
- In the direction of **AI governance**, there is a lack of concrete policy suggestions.
- The direction of **corrigibility** focuses on an online correction of algorithms. The so-called "corrigible reasoning" deals with the design of agents, so they update their reasoning and do not have the benefits of escaping or manipulating something (Soares and Fallenstein, 2017).
- The direction of **safe exploration and distributional shift** handles detecting and adjusting behaviors of agents to prevent mistakes (Amodei et al., 2016).
- **Adversarial robustness** describes the need to minimize the success rates of adversarial attacks.
- **AI Ethics** differ between cultures and evolve over time (Awad et al., 2018; Hagerty and Rubinov, 2019). There is a need to research moral preferences.

Furthermore, the European Parliament, Council, and the European Economic and Social Committee review the safety implications of legislation concerning AI. For instance, the legislation should contain requirements to address the risks of faulty data input at the design phase and processes to ensure quality. In addition, the report states that a risk assessment is necessary not only before entering a product into a market but also before important changes during the product life-cycle (European Commission, 2020).

Besides Juric et al. (2020) and the report of multiple European institutions, we especially found several blog entries dealing with literature reviews on AI Safety (c.f. Dawson, 2017; Krakovna, 2021). Hence, there is a need for further peer-reviewed literature reviews, specifically in areas that are not covered in Juric et al. (2020) and the EU report.

4. Systematic literature review

This chapter presents the applied, systematic process of collecting and reviewing the scientific works on the topic. Besides the approach, the data collected during the review and a summary are provided.

TABLE 2 Methodology for the systematic literature review (adapted of Brereton et al., 2007; Kitchenham and Charters, 2007; Schumann et al., 2020.

Chapter	Phase	Activity	Description
4.1.1	Planning	Select database	Select based on research questions and keywords
4.1.2	Conducting	Forward search and cluster	Search in database with search phrases and cluster
4.1.3	Conducting	Backward search	Search in references of publications
4.1.4	Conducting	Select relevant publications	Select based on criteria
5	Documenting	Document approaches	Document classical approaches
6	Documenting	Document approaches	Document new approaches

4.1. Research methodology and its application

In order to conduct a valuable review of the existing literature and cover the addressed topic, a systematic approach is required. The literature review is based on the process described in Brereton et al. (2007), Kitchenham and Charters (2007), Liberati et al. (2009), and Schumann et al. (2020). In the planning phase, the database is created, based on research questions, one example and keywords. The conducting phase consists of a forward search which establishes the reference list. The following backward search investigates this reference list. The documenting phase in chapter 5 and 6 focuses on the identified, relevant publications.

The following Table 2 provides an overview of six steps which are part of the three research phases of planning, conducting and documenting introduced in Brereton et al. (2007).

Figure 2 illustrates the literature search based on the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) checklist (c.f. Liberati et al., 2009) with the number of relevant publications in each step of the search process.

4.1.1. Planning phase—Select database

4.1.1.1. Research questions

The methodology starts by formulating the research questions, which are derived from the problem of applying unsafe systems in unknown environments. The answers to the following questions are sought in the context of the example described in the subsection "Example" following "Keywords":

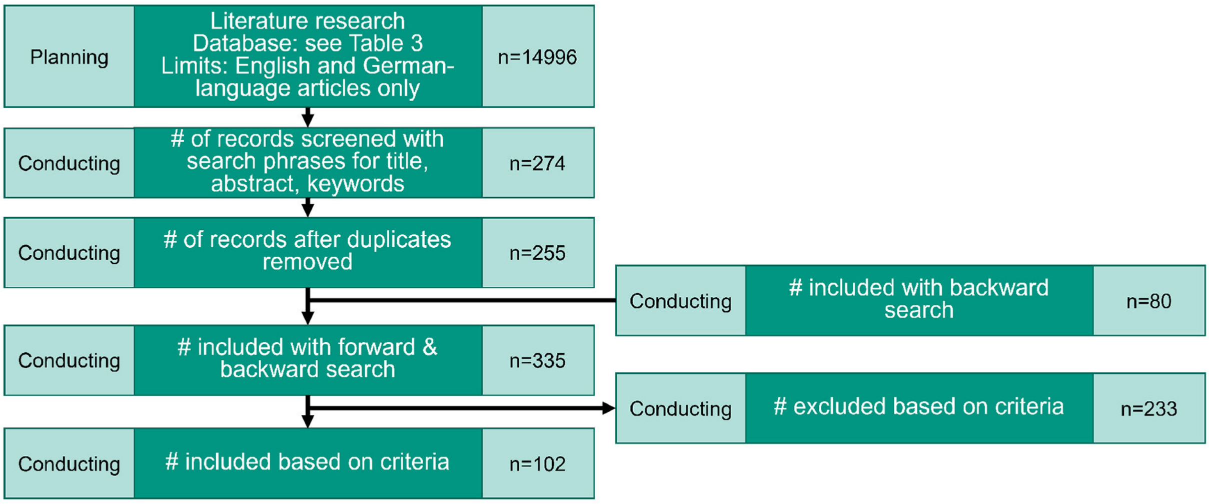
Planning
Literature research
Database: see Table 3
Limits: English and German-language articles only
n=14996
Conducting
# of records screened with search phrases for title, abstract, keywords
n=274
Conducting
# of records after duplicates removed
n=255
Conducting
# included with backward search
n=80
Conducting
# included with forward & backward search
n=335
Conducting
# excluded based on criteria
n=233
Conducting
# included based on criteria
n=102

FIGURE 2
Overview of the number of publications in each activity based on PRISMA (c.f. Liberati et al., 2009).

FIGURE 3
Keywords.

- What have been the trends in AI over the last 10 years?
- Which are the problem classes for AI Safety in the context of highly automated driving?
- Do suitable validation approaches exist in the identified problem classes?
- Can complex decision-making strategies be learned and tested?

These various questions can be summarized into one guiding question: **Which approaches and methods can be applied for the testing, verification, and validation of AI systems for highly automated driving?**

4.1.1.2. Keywords

In the beginning, core keywords were identified (see Figure 3). The keywords were substantiated based on the research need and example, as well as on the results in the forward search.

Due to the large number of contributions we have found in this way, we have decided to narrow the focus and exclude the following topics.

- Improvement of systems through AI Algorithms, e.g., sensor fusion.
- AI and Ethics.
- The Security of AI and cyber-physical systems.

4.1.1.3. Example

In order to narrow down the focus of the research, this section describes a reference example.

Another aspect of understanding the challenges *AI Safety* faces is to investigate the intended application in which AI can be used. The question here is whether problems can be derived from the applications that AI can better solve.

In general, the use case is intended to consider autonomous driving. Autonomous driving is the autonomous arrival at a given destination without the intervention of a human being. For this purpose, the vehicle must be able to orient itself in the environment and make independent decisions.

In order to be able to make decisions, the vehicle must be able to identify objects in its environment independently and derive specific actions from them. According to SAE J3016, this corresponds to SAE level 4, i.e., a safety-critical system (SAE International, 2021).

A vehicle is a complex system with multiple subsystems involved in achieving its functionalities from a safety perspective. For instance, the navigation system has the task of suggesting a route to the destination from the vehicle's current position. A control computer is then responsible for controlling the active steering system from the suggested path

TABLE 3 Overview of the database sorted by number of publications.

Type	Publication	Timeframe	Count
Journals	IEEE Transactions on Intelligent Transportation Systems	2015–2021	2,400
	Reliability Engineering and System Safety	2015–2021	790
	Journal of Artificial Intelligence Research	2015–2021	360
	Artificial Intelligence for Engineering Design, Analysis and Manufacturings	2015–2021	303
	German Journal of Artificial Intelligence	2019–2021	174
	Frontiers in Artificial Intelligence and Applications	2015–2021	140
Conferences	IEEE Conference on Computer Vision and Pattern Recognition	2015–2021	5,817
	International Conference on Software Engineering	2015–2021	4,742
	International Conference on Artificial Intelligence and Advanced Manufacturing	2019-2021[a]	177
Further Publications	Workshops: WAISE—Workshop on Artificial Intelligence Safety Engineering, AI Safety Workshop	2019–2021	70
	Standards: ISO26262-(1-12:2018), ISO/PAS 21448:2019, ASAM OpenX, ANSI/UL4600-2020	2018–2021	18
	Blog articles[b]	–	5

[a] Available to the authors on ACM Digital Library.

[b] Blog articles represent personal opinion without any peer review process. The following blog articles were considered: Ortega et al. (2018), Burton (2021), Gauerhof et al. (2021), Krakovna (2021), and Faculty.

so that the vehicle reaches its destination. A situation awareness system derives the current situation of the vehicle and the environment from various sensor information. In a hazardous situation, the vehicle must be controlled and stopped by a brake assistant, for example, so that no severe damage is caused. The interaction of various independently acting systems enables the realization of an autonomously driving vehicle.

The vehicle operates in an Operational Design Domain (e.g., public road) and interacts with other systems like traffic participants and infrastructure. This so-called System of Systems can contain AI systems. The typical challenges of System of Systems like different life cycles, operational and managerial independence, and many stakeholders with sometimes conflicting objectives need to be addressed.

In addition, an autonomous vehicle is embedded in other systems. If the position and speed are continuously communicated to a traffic control system, the current traffic situation can be created based on this data. If a forecast is also provided, the navigation system can determine and consider alternative routes. Moreover, newly learned strategies for successfully evaluating and reacting to a situation can be transferred to a fleet of autonomous vehicles. To support the situation representation system, access to external sensors may improve the identification of objects and the system's reaction.

Thus, an autonomous driving car is a cyber-physical system that interacts with other systems in a highly complex System of Systems.

4.1.1.4. Database

With import functions on high-ranked journals and identified conferences as well as other publications, a literature basis for the forward search was created. The authors identified that not only journals and conferences but also internet documents contain new and relevant information, which should be considered (see Table 3; c.f. Schumann et al., 2020). The total number of researched publications is 14,996.

The literature management tool Citavi allows to collect the publications. Afterwards, an analysis is conducted in the tool MAXQDA (VERBI Software, 2021).

4.1.2. Conducting phase—Forward search and cluster

In the following, the results of the forward literature search are described with metrics. During the forward search, nine main categories in the field of AI Safety were identified (see Figure 4).

Figure 5 shows how often the selected keywords in this contribution are used in the publications over the years. There is high volatility in the usage of certain keywords like robustness. In conclusion, there is no clear tendency. Possible causes are many safety-related documents in 2016 and the tendency to use more specific keywords.

As Figure 6 points out, the literature research resulted in more than 55 relevant documents in the area in focus. As a result, the authors conclude that the amount of publications is sufficient for the forward search phase.

In conclusion, over the past 6 years, there has been a shift of topics more to safety, vehicle(s), and monitoring. Furthermore, the high percentage of certain keywords indicates a good selection of the publications.

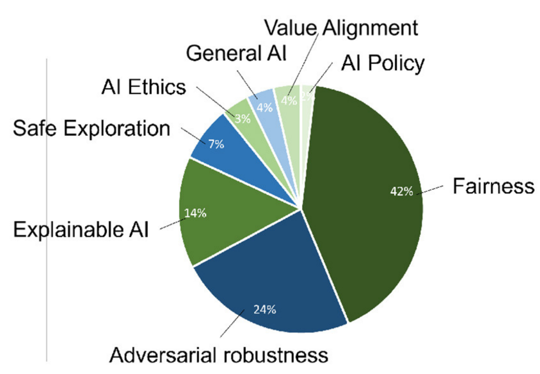

FIGURE 4
Relative occurrences of categories in database.

4.1.3. Conducting phase—Backward search

The backward search in references of identified publications result in 80 more publications to consider (see Figure 2). Based on the results of the backward search, the list of contributions was divided into the two areas, "established approaches and methods" and "new approaches and methods" (see chapters 5 and 6). The two areas were defined to give a better overview and address the possible use of established approaches and methods.

4.1.4. Conducting phase—Select relevant publications

The final selection of relevant publications are based on the following criteria:

- AI Safety-related

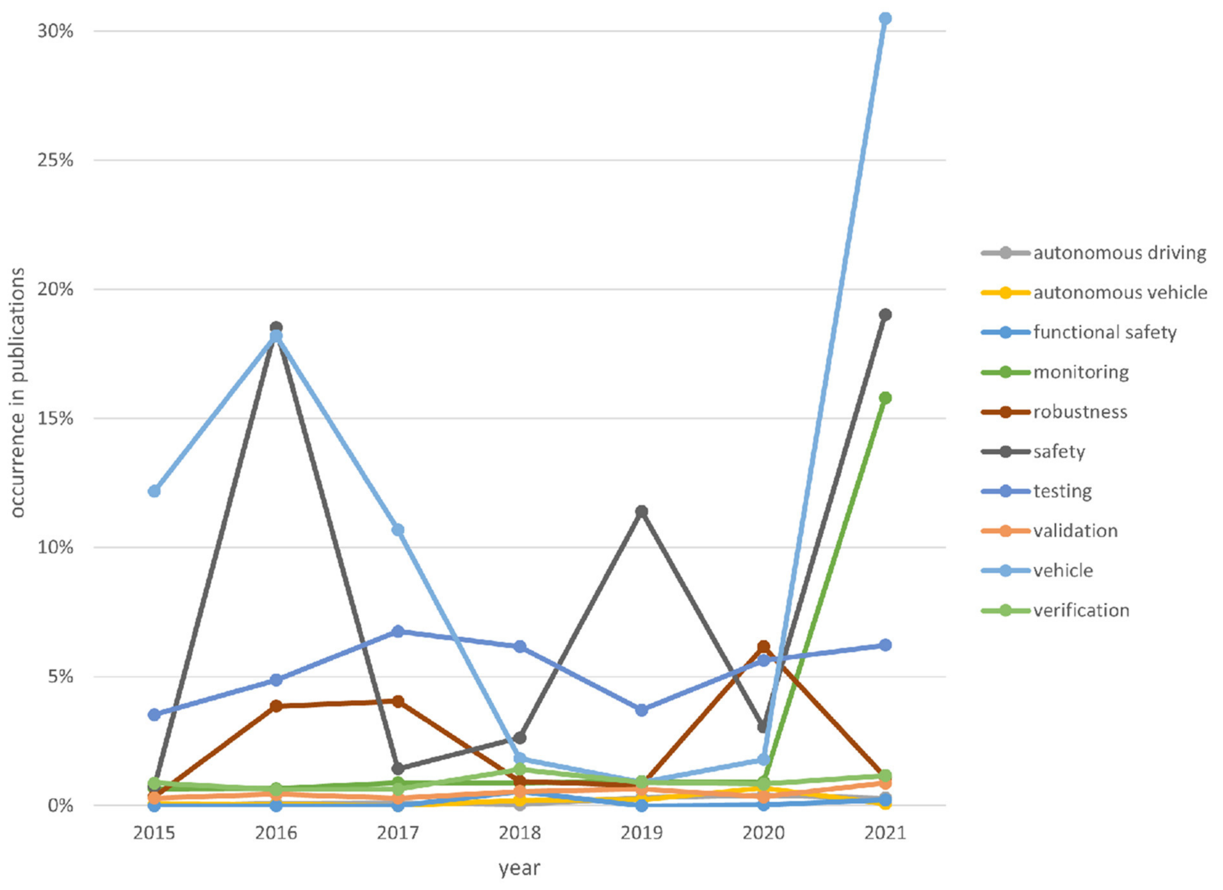

FIGURE 5
Occurrences of keywords in publications.

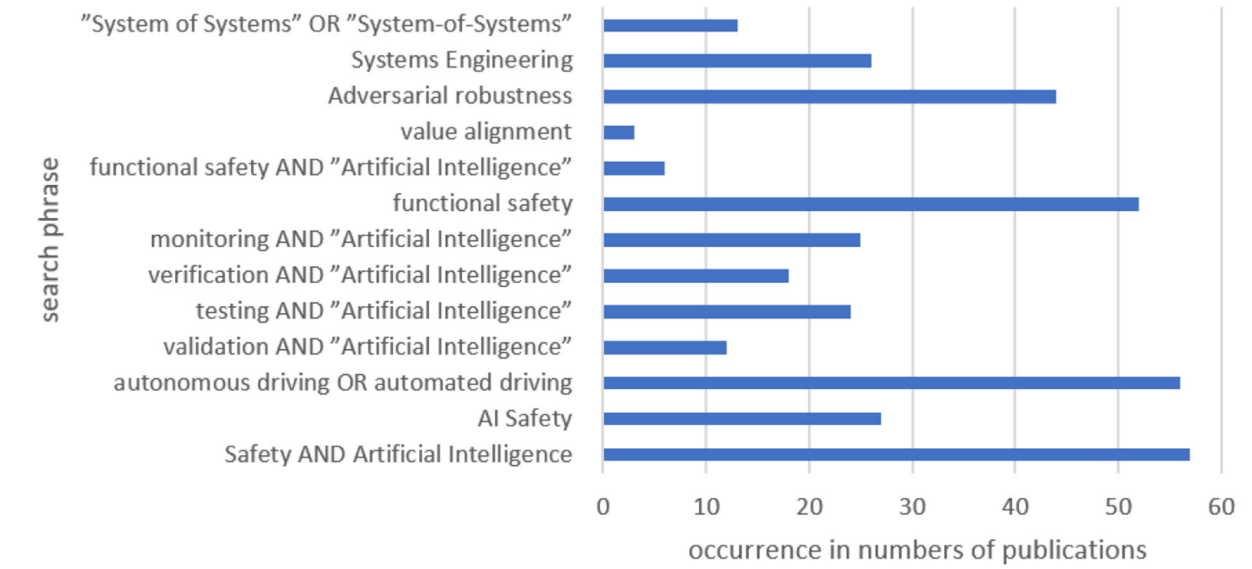

FIGURE 6
Occurrences of search phrases in publications.

- Selection of high-ranked journals and conferences. Selection of blogs, norms, and standards (see Table 3)
- Identified with search phrases or further references identified with backward search (see Section 4.1.2)
- Suitable for the example of highly automated driving (see example in Section 4.1.1).

Publications related to the criteria "non-English and non-German articles" and the keywords "ethics, security, improvement of system(s)" are excluded to limit the scope.

According to the results of the backward search, the following particularly relevant topics emerged: robustness (6, 27) (read: number of publications in 2015, number of publications in 2021), value alignment (0, 2) and validation (10, 15), verification (16, 30), testing (72, 105).

5. Classical approaches

Classical or established approaches describe existing approaches for non-automated and automated systems. For instance, the neural network's relevant approach to neuron coverage resembles the traditional code coverage testing in computer science. Hence, it is categorized as a classical approach. As a result, classic or established approaches are already published standards and established concepts in different research fields over more than 10 years. In the first section, classical approaches are summarized based on model approaches.

5.1. Modeling approaches

Different verification models exist to systematically and formally verify statements of an argument. The field of argumentation theory deals with arguments for logic and rhetoric. An example of a general statement is the claim stated in Figure 7. The already 1958 published Toulmin model with additional content in a new edition (Toulmin, 2003) is shown in Figure 7. This classical approach can handle autonomous learning systems with logic (Collopy et al., 2020). Starting from a claim, one or multiple warrants describe the guarantees for this claim. These can be rebuttals or rely on evidence.

Further extensions are done by Hirata and Nadjm-Tehrani (2019) with a combination of Goal Structuring Notation with the Systems Theoretic Process Analysis to support the safety claim.

In the field of Neural Networks, Kurd's Neural Network Development Model can be applied. It integrates hazard analysis in the development of the neural network's knowledge and deals with neural networks developed specifically for safety-critical use (Kurd and Kelly, 2003).

Shalev et al. introduce a formal model of safe and scalable self-driving cars. The contribution of the paper is two-fold:

Firstly, a mathematical model called "Responsibility Sensitive Safety" (RSS) formalizes an interpretation of "duty and care." It is designed to achieve the following three goals:

- Its application complies with how humans interpret the law of duty of care.

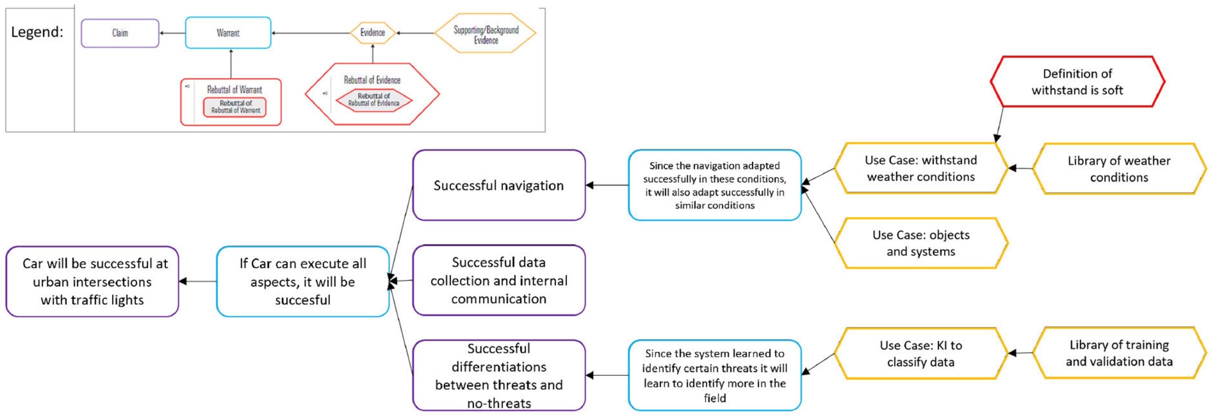
Legend:
Claim
Warrant
Evidence
Supporting/Background Evidence
Rebuttal of Warrant
Rebuttal of Rebuttal of Warrent
Rebuttal of Evidence
Rebuttal of Rebuttal of Evidence
Definition of withstand is soft
Successful navigation
Since the navigation adapted successfully in these conditions, it will also adapt successfully in similar conditions
Use Case: withstand weather conditions
Library of weather conditions
Use Case: objects and systems
Car will be successful at urban intersections with traffic lights
If Car can execute all aspects, it will be succesful
Successful data collection and internal communication
Successful differentiations between threats and no-threats
Since the system learned to identify certain threats it will learn to identify more in the field
Use Case: KI to classify data
Library of training and validation data

FIGURE 7
Example classified according to the Toulmin method—extract (c.f. Toulmin, 2003).

- The interpretation leads to a useful driving policy, i.e., it should lead to an agile driving policy rather than overly-defensive driving.
- The interpretation is efficiently verifiable so it can be thoroughly proven that a self-driving car correctly applies the interpretation of the law.

Secondly, a semantic language consists of units, measurements, action space, and specifications. The language is used to plan, sense, and actuate an autonomous vehicle. The authors state in particular that the model guarantees (from a planning perspective) that there will be no accidents caused by the autonomous vehicle (Shalev-Shwartz et al., 2018).

Further safety argumentation is realized in Burton et al. (2019) and Schwalbe et al. (2020). Burton et al. (2019) developed an approach for building confidence arguments. These arguments are used for the evaluation of performance in machine learning. The approach was applied for the evaluation of pedestrian recognition.

5.2. Norms, standards, and roadmaps

Multiple norms and standards currently address the topic of AI Safety. Furthermore, the domain focus of this contribution is on highly automated driving. For a better overview, the contributions are chronologically introduced. They focus on approaches as well as challenges, guiding questions, requirements, and classifications in the context of AI Safety.

2018: For automotive safety, the ISO 26262:2018 addresses functional safety. The norm categorizes the risk by Automotive Safety Integrity Levels (ASIL) to differentiate the proposed measures of handling the risk levels. Inductive methods like the Failure Mode and Effects Analysis (FMEA) have a bottom-up approach to start with the occurred effect first, subsequently deriving the effect's causes. In addition, deductive methods like the Fault Tree Analysis (FTA) use a top-down approach (Salay et al., 2017; ISO Central Secretary, 2018b, p. 11). These result in safety considerations starting from the system and refining it to components with, for instance, their probabilities of default. With the FTA, a probability of default on the system level can be derived.

2019: In April 2019, the European High-Level Expert Group on Artificial Intelligence (HLEG-AI) (HLEG, 2019; Independent High-Level Expert Group on Artificial Intelligence, 2020) published Ethical, Legal, and Technical "Key Requirements" for reliable AI systems. The seven requirements include:

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-discrimination, and Fairness
- Societal and Environmental Wellbeing
- Accountability.

In a report considering trust in human centered AI, it is stated to realize safety, a fallback plan and a proactive testing of safety measures is necessary. Moreover, safety measures should depend on the risk posed by AI systems. (HLEG, 2019; Independent High-Level Expert Group on Artificial Intelligence, 2020).

In the white paper titled "SAFETY FIRST FOR AUTOMATED DRIVING," multiple automotive companies address the issue of nonexistent solutions in the topics of automated driving, e.g., safety assurance of AI. They argue that safety-related use cases need to be analyzed for different safety level assurances. Other topics of interest are functional descriptions, assessment, development process, verification, validation, experiments, safety analyses, and safe design (Wood, 2019).

The roadmap SafeTRANS subdivides ten categories for future human-machine systems: integrity and certification, cooperation, context, strength, responsibility, and reflection. For instance, the categories autonomy is subdivided in maneuver, mission, collaborative, and autopoietic autonomy (SafeTRANS, 2019a,b).

Diverse norms rank and categorize future areas of interest according to categories like IT-safety and security (SafeTRANS, 2019a).

2020: In 2020, the HLEG-AI presented its final assessment list for Trustworthy Artificial Intelligence. (European Commission, 2020) In total, the group identified four ethical principles and seven requirements which companies can follow to achieve trustworthy AI. For the general safety of AI, the AI HLEG derive five key questions:

- "Did you define the AI system's risks, metrics, and risk levels in each specific use case?"
- "Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences?"
- "Did you assess the dependency of a critical AI system's decisions on its stable and reliable behavior?"
- "Did you plan fault tolerance *via*, e.g., a duplicated system or another parallel system (AI-based or 'conventional')?"
- "Did you develop a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness and safety?"

In contrast to the specific autonomous focus of some norms and standards, the UL4600 is the standard for safety for evaluating autonomous products. The standard gives mandatory and recommended advice in five subcategories for each topic. For machine learning, the standard gives the following six criteria to be considered: acceptable capabilities (1), acceptable performance (2), acceptable data (3), robust to data variations (4), the post-development does not comprise the safety (5), safety also for every other AI not considered in machine learning (Underwriters Laboratories, 2020).

2021: Other organizations state questions to address the "future" topic of safety. The European initiatives OSS.5 hosts an event for system safety for SAE level 4 and 5 automated vehicles. On its website, the questions are risen (Tomorrows Business GmbH, 2022):

- How to formulate a continuous safety case in the field of autonomous driving?
- How to integrate safety of functional operations in the field of AI, ML, and Deep Neural Networks?
- How to ensure that deep learning-based systems are safe?

The International Telecommunications Union focuses on "AI for Autonomous and Assisted Driving." The aim is to develop performance measurement standards for AI that control self-driving vehicles. In particular, a data protocol for Safe AI is in development (International Telecommunication Union, 2022).

5.3. Classical approaches and methods of non-norm and non-standard publications

5.3.1. Verification

Cheng et al. (2017) propose extensions for existing safety standards for the usage of neural networks. They extend the aspects "Implementation understandability," "Implementation correctness," and "Specification validity" from the existing standard toward safety certification of neural networks. A concrete use case is presented, and a reference to a NASA report covering the topic in the aeronautic area is given (Cheng et al., 2017).

5.3.2. Validation

In the field of validation, Ebert and Weyrich (2019) summarize mainly non-data-driven AI. According to the authors, the validation technologies for autonomous systems can be subdivided into white-box/black-box validation strategy and manual or automatic validation handling. For AI purposes, black-box validation strategies, in particular, should be focused on, as the authors believe AI has a black-box character. As a result of this, the following methods are evaluated in the contribution (Ebert and Weyrich, 2019):

- Experiments and empirical test strategies
- Specific quality requirements tests, for instance, penetration testing, and fuzzing
- Brute-force usage in the real world while running realistic scenarios
- Intelligent validation, for instance, cognitive and AI testing.

5.3.3. Testing

One approach for testing neural networks for autonomous driving with synthetic data is described by the authors (Dreossi et al., 2017). With synthetic data, the authors could test the CNN[4] to detect cars. This approach can be seen as a classical way of handling the safety analysis by modeling the environment

[4] CNNs, short for convolutional neural networks, refer to special artificial neural networks, which are used particularly frequently in the field of image processing. For a brief introduction to the topic, see for example O'Shea and Nash (2015) and Albawi et al. (2017).

with synthetic data to train the CNN. Combined with further approaches to analyze the robustness of CNNs, this may help generate enough training data. Further research on synthesizing sensor data is done in Yang et al. (2020).

Koopmann introduces safety cases for SAE level 3 testing. He argues that better autonomy results in more challenging situations. The test platform with different actors and systems involved might be partly applicable for higher levels of automated vehicles (Törngren, 2019).

Other research focuses on search algorithms to find (test) data generation in the context of autonomous driving (c.f. Han et al., 2021).

5.4. Extract of other classical approaches

An extract of other classical approaches in three different fields is shown in the following.

In the context of **neurological research**, Sotala (2015) researched safe AI with concept learning methods from humans. He reviews multiple approaches and extracts basic steps for concept learning.

The authors' (Page et al., 2018) discuss issues and risks (such as malfunction, malicious attacks, mismatch of objectives) appearing in the usage of AI systems. For this purpose applications of **agent-based systems** are considered. Some of the accompanying risks (and potential strategies for mitigating them) are discussed. These include the change of objectives and the exploration of the environment. This is especially beneficial if the AI has unlimited access to all environment variables. Due to security reasons, this might not be realizable (Page et al., 2018).

In the context of **System of Systems**, the topics of safety and Artificial Intelligence are important. For example, the word "Artificial Intelligence" is used 25 times and safety 121 times in the book "Disciplinary Convergence in Systems Engineering Research." It is pointed out that a socially acceptable degree of reliability and safety of highly autonomous vehicles can not be assured by treating the vehicle solely as a software system (Koopman and Wagner, 2016). More likely, the vehicles must be seen as part of a System of Systems (Boehm et al., 2018).

5.5. Current gap in classical approaches and the resulting need for new approaches

In summary, the classical approaches focus on established approaches in multiple research fields. Classical approaches focus on SAE levels 1–3 and focus on known environments and applications. However, considering higher SAE levels, new approaches are necessary for AI Safety. Moreover, the safety consideration of highly connected AI systems can be further considered. In order to deal with these new challenges of highly autonomous systems like vehicles, the next chapter depicts new approaches for validation, verification, and testing.

6. New approaches and methods

Due to the identified gap, new methods are needed. In addition, there is the challenge of autonomous driving (SAE 4) as an application of AI in specification, design, and implementation. In the following, new approaches are described for verification, validation, and testing. Exemplary Adversarial Robustness and Value Alignment are considered in more detail.

6.1. Verification, validation, and testing

Many AI systems—for example, ML systems—differ considerably from classical software solutions. Classical software is characterized due to a set of instructions translated into program code by a developer. Applying these instructions to input data gives the output, i.e., it is evident how the output data depends on the input data. In contrast, ML approaches try to extract inference knowledge from so-called training data. If the ML system is based on a neural network, the acquired knowledge is represented in terms of a parameter-dependent model—the neural network.

Neural networks are created by the concatenation of a typically large number of mathematical operations and can schematically be represented in layers consisting of neurons (see, for example, Abdi, 1994). ANNs consisting of multiple layers are called deep neural networks (DNN) (see Montavon et al., 2018). The deeper a network, i.e., the more layers a DNN possesses, the more complex the input and output variables of the network are linked with each other. Consequently, if only the effect of a DNN on given input data is known, i.e., only the corresponding output can be observed, it is difficult to infer the exact functioning of the DNN. For this reason ANNs are often characterized as a black box. According to that, a classical safety analysis is not applicable since a complete understanding of the system's functionality would be required. Additionally, due to the lack of an existing instruction set, traditional validation, and verification methods are not suitable for learning systems. Thus, as requested in Droegemeier et al. (2019), new methods for safety assessment and, in particular, for verification and validation are required.

6.1.1. Verification—adversarial robustness

In recent years, a requirement that has gained importance is robustness against so-called adversarial attacks. It refers to the vulnerability of neural network-based classifiers with respect to small perturbations in the input data. As shown in

Moosavi-Dezfooli et al. (2016) or Kong and Liu (2019) at any fixed input sample, for humans, imperceptible perturbations can be constructed such that the resulting perturbed input is misclassified. Inputs of this kind are termed adversarial examples. A taxonomy of adversarial examples and a review of different methods for their generation and as well on countermeasures against adversarial attacks can be found in Yuan et al. (2019). Two trends can be identified for improving adversarial robustness. There are, on the one hand, approaches focusing on the enrichment of the training data-set with a wide variety of adversarial examples (see for example Szegedy et al., 2013; Tramèr et al., 2017; Song et al., 2018) and on the other hand methods based on the adequate choice of the underlying cost function (see Goodfellow et al., 2014; Madry et al., 2017). As addressed in Tsipras et al. (2018), Su et al. (2018) increasing the adversarial robustness of a classifier may negatively affect its accuracy. In contrast to this, however, in Mao et al. (2019) and Stutz et al. (2019), it has been shown that it is, in fact, possible to develop robust and accurate neural network-based classifiers.

Besides developing training strategies to reduce the vulnerability of ML systems against adversarial attacks, research is being conducted to find formal guarantees on its robustness—see for example Hein and Andriushchenko (2017). Formal verification methods, in general, seek to prove that desired properties are satisfied using mathematical reasoning. Even though ANN is a well-defined concatenation of mathematical operations, a formal analysis of its functionality is usually not suitable because of its size and complexity. Hence, citing Katz et al. (2017a), only automatic verification techniques are needed. According to Katz et al. (2017a) again, it can be shown that this problem is nondeterministic polynomial-time complete (or short NP-complete—see for example Goldreich, 2010) and thus difficult to solve. Progress was made in this respect for ANNs based upon ReLU[5] activation functions—see once more Ehlers (2017), Katz et al. (2017a), or Katz et al. (2017b), Singla and Feizi (2019).

6.1.2. Validation—value alignment

For the validation of ML systems, the area of value alignment is of particular interest and importance. It tackles the issue of developing a system following the intentions of its developer (see Soares et al., 2015; Taylor et al., 2016 or as well Hubinger et al., 2019). To illustrate this issue, consider the fictional cleaning robot problem presented in Amodei et al. (2016). In order to keep an office free from messes, a cleaning robot relying on a ML controller trained by means of reinforcement learning (RL) techniques shall be used. Assume that the training is designed so that the robot gets rewarded only if it is not perceiving any messes in the office. Then the most profitable strategy in terms of reward maximization may be obtained if the robot deactivates its perception system, which ensures that it will not find any messes at all. This, of course, gives a solution that does not solve the actual problem of keeping the office clean. This example emphasizes that training a ML system may result in optimal solutions to the underlying optimization problem but, in general, does not consider other aspects. While harmless in the context of such an example, this phenomenon—called reward hacking (see Amodei et al., 2016)—may have serious implications in other circumstances like the control of self-driving cars, which implies the special relevance of this topic in the given scope. To this end, in Taylor et al. (2016) two main issues are pointed out: specification of the right kind of objective function and the design of ML systems which avoid undesirable behavior even if not perfectly aligned with its developer's intentions. The authors designate and review eight research directions that, according to their conclusions, may be beneficial for the development of reliable and safe learning systems.

The danger that the incorrect choice of the cost function can lead to unexpected and undesirable side effects is also addressed in Amodei et al. (2016). Therein specific attention is paid to the role of RL in the context of value alignment. RL is based on regular interactions between the ML system and its environment, suggesting that this technique can be used as a valid strategy to solve the value alignment problem. However, this assumption is challenged by the possible occurrence of reward hacking, as the cleaning robot example emphasized. A promising extension of RL for tackling the value alignment problem is inverse reinforcement learning (IRL)—citing (Hadfield-Menell et al., 2016) IRL is certainly relevant to the value alignment problem. In contrast to RL, where an optimal control is derived solely by the interaction of the ML system with its environment given a fixed, predefined reward function, approaches to IRL are based, in simple terms, on the idea of mimicking control strategies are considered optimal. More precisely, by observing a system that acts according to an optimal control strategy, one aims to derive the underlying reward function, which is then used to train the ML system. For a more detailed elaboration, see Ng et al. (2000), Hadfield-Menell et al. (2016), or as well Finn et al. (2016).

6.1.3. Testing

Freeman (2020) provides a list of themes that may serve as a starting point for test and evaluation methodologies of data-driven AI systems. In particular, a risk-based test approach is recommended, and the usage of metrics is highly encouraged. Concerning the latter already existing concepts may properly extended or adjusted—see Cheng et al. (2018) and again Freeman (2020). Additionally Skias (2006) suggests tackling

[5] ReLU, short for Rectified Linear Unit refers to the function $x \in \mathrm{R} \mapsto \max(0, x)$. It is a popular choice for activation functions of artificial neural networks—see for example Goodfellow et al. (2016) and Sharma et al. (2017).

during the assessment of a learning system, among others, the following topics:

- Has the correct data been learned, or has been learning something else closely related to the data?
- Did the training procedure give a global optimum or only a local one?
- How does the system react to unseen data or edge cases?

As for non-data-driven systems, any test concept of ML systems has to comprise criteria that allow proving a statement by providing sufficient evidence in favor of this statement. As a procedural approach for this purpose, one may consider again the generic Toulmin-method or, for safety-relevant questions, the approach presented in Kurd and Kelly (2003).

6.2. Development model

The increasing usage of applications with machine learning components calls, following Serban et al. (2020), for mature engineering techniques which ensure these are built robustly. As pointed out in Arpteg et al. (2018), there are fundamental differences between developing systems comprising ML solutions compared to traditional software systems. Thus, there is, as pointed out also in Ammar et al. (2006), the need to extend classical software development models like the waterfall methodology or the spiral model. In this regard, the authors of Ammar et al. (2006) discuss three different approaches:

- Common ML system development model: Iterative cycle consists of design, training, and testing stages.
- Rodvold's model: Incorporating nested development loops and containing elements from the waterfall and spiral model.
- Kurd's model: Comprising hazard analysis and addressed in particular for the development of ML systems in safety-critical applications.

An insight into the ML workflow of Microsoft workgroups is provided in Amershi et al. (2019). It consists of nine stages, which are pooled into two groups: data-oriented and model-oriented. The first group deals with data acquisition and data cleaning, while the second group is dedicated, among others, to training, evaluation, deployment, and monitoring. According to findings in Lorenzoni et al. (2021), a very recent work giving a systematic literature review on the development of machine learning solutions, this process is the most comprehensive and accepted one in the literature.

A similar process for the development of ML solutions is provided in Hesenius et al. (2019). In their proposal of a development process, the authors define phases and roles that are to be assigned to members of the development team and are interwoven with the flow of the process. They name here domain expert, data scientist, data domain expert, and software engineer. According to this process, in the first phase, data scientists have to decide whether an ML solution is suitable for the problem. The first model is implemented only after exploring the available data (phase 2), and requirements are defined (phase 3). The last two phases of the process deal with the integration and operation of the developed components. Even though the authors propose an agile development process, they emphasize that the method described can also be applied in a more rigorous process.

7. Discussion

Many of the topics discussed in this article can be considered apart from AI Safety in light of other research areas. Specifically, this concerns value alignment and adversarial robustness. While the former can be assigned to the area of AI ethics (see Vakkuri et al., 2019; Kazim and Koshiyama, 2021), the latter is according to Carlini et al. (2019) a security-relevant topic. The importance of security in state-of-the-art vehicle development is discussed, for example, in Schwarzl et al. (2021). The authors discuss the impact of security risks on safety and outline safety and security co-analysis and co-design methods in autonomous driving. From this article, it can be deduced that security aspects must also be considered for a fully comprehensive safety analysis.

A major difficulty in validating and verifying data-driven approaches that should not be overlooked is that all statements about the performance of such an approach are of statistical nature. For example, all that can be said about an ML system for traffic sign recognition is that it correctly recognizes a certain percentage of all signs belonging to a test set. Even if the recognition rate on this test set is one hundred percent, it cannot be assumed that the system works appropriately outside of it, i.e., without false recognition. As was pointed out in Section 6.1.1, even minor disturbances in the input data can lead to misclassifications. Consequently, the quantification of the uncertainty or, more precisely, the specification of an upper bound for the uncertainty with which a ML system makes decisions is highly necessary.

The topic of uncertainty quantification (UQ) is not a unknown one in the science community. The increased use of ML solutions in decision-making has brought this area back to focus in recent years. An overview of the topic is provided for example in Begoli et al. (2019), Abdar et al. (2021), Seuß (2021), or Psaros et al. (2022). In regard to UQ, or more general, in regard to management with uncertainty, we would also like to draw attention on the fuzzy logic approach. According to Zadeh (1983), fuzzy logic or fuzzy reasoning provides an alternative approach for the management of uncertainty. It refers to a multi-valued logic allowing to introduce values between two extreme characteristics like 0 and 1 or true and false. According to Hellmann (2001) and Zadeh (2008), this approach applies human-like thinking in the programming of computers and provides the capability to reason and make rational decisions in an environment of imperfect information. In the context of the (automated) control of a vehicle, fuzzy logic allows for example to formulate and model mathematically notions like steer, steer sharp, brake, or brake hard. For more in-depth information on this topic, its implementation and possible applications, especially in the field of automated driving, we refer to Lee (1990), Passino et al. (1998), Peri and Simon (2005), Dey et al. (2016), and Masmoudi et al. (2016).

8. Open challenges and future trends

There is an increasing amount of publications in the area of AI Safety. The authors think this trend will intensify with even more publications in the future. There are promising approaches, especially in the investigated areas of validation/verification/testing. Splitting up the safety assessment of AI systems into the separate tasks of validation, verification and testing seems to the authors as a promising point of reference for future development in the field of AI Safety. The topics of validation/verification/testing will need to be addressed from general perspectives as well as for specific problems.

Several standards and norms in the area of AI Safety of highly autonomous vehicles emphasize the high relevance of the topic. However, as mentioned in Salay et al. (2017), standards like ISO26262 would need to extend their notion of "hazard" toward cases of AI interaction with humans. A first step in this direction can be expected from the publication of ISO/IEC DTR 5469734 Artificial intelligence—Functional safety and AI systems, which is still under development. The authors believe that the introduction of standards or the extension of existing standards to AI systems is essential and one of the prerequisites for the safe application of AI systems in control of safety-critical systems.

Systems consisting of AI are getting increasingly complex. This results in big systems with many interrelations and stakeholders involved. Terms like Advanced Systems, Cyber-Physical Systems and Product-Service Systems can contain AI and need to be safely designed. Hence, the scope of AI Safety should not be limited to one kind of system and authors should consider connectives between systems.

Besides the system's autonomy coming from AI within systems, most systems are socio-technical. Therefore, the interaction with humans is an important topic not covered in this publication. Hereby, multiple standards and norms address this topic and give an outlook (c.f. ISO Central Secretary, 2018a; HLEG, 2019; Koopman et al., 2019b).

The automotive infrastructure is critical due to its many challenging demands for safe and reliable systems. While the main focus is on ground transportation, there are influences and connections to other areas such as aerospace by employing drones or the energy sector by connecting vehicles to a (smart) charging grid. Hence, safety standards and regulations have to be aligned. Classical approaches in other fields like aerospace can be taken into account and can have substantial benefits in the automotive system development. In the context of unknown environments, coping with uncertainties is a crucial task. As already mentioned in Section 7 approaches based on fuzzy logic may be helpful in this regard.

9. Conclusion and limitations

Due to the rising number of machine learning solutions and more general AI solutions for the partial or full control of safety-critical systems, the field of AI Safety becomes increasingly important. For the purpose of providing an overview of this extensive research area, a systematic literature review focusing on the field of highly automated driving was conducted. It appears that the topic of AI Safety has become more important over the last years. Despite fluctuations, we identified a trend of publications considering our keywords increasing from 368 (2017), 2,778 (2018) to 2,844 in the year 2021. Furthermore, the review shows that the term AI Safety is only found 12 times. Other search phrases like "AI Testing" or "Safety AND Artificial Intelligence" are mentioned more often. Therefore, it seems that the term "AI Safety" is not yet well-established. The cumulative research question of this paper examines the approaches and methods which can be applied for the verification, validation and testing of AI systems for highly automated driving. According to the findings of this literature review, we identified two major branches comprising approaches and methods for the safety assessment of AI systems: classical approaches, such as the Toulmin method (Section 5) and newly developed approaches (Section 6) tailored to ML use cases, such as value alignment. In addition, the literature review revealed that the aerospace industry has already been facing challenging topics like verification and validation of AI systems for a considerable amount of time—see for example Bhattacharyya et al. (2015) or Underwriters Laboratories (2020). This emphasizes that for the development of safe AI systems in the automotive sector, findings from the aerospace sector should be taken into account and may be used as a guidance. Furthermore, we conclude that approaches for the safety assessment of AI systems can be considered in a general framework superior to the specific use case.

Limitations

The findings of this work have to be seen in the light of some limitations, which are listed and discussed briefly in the following:

- Six high-ranked journals and three conferences are considered. Furthermore, norms, standards, workshops and blog entries are included, since the topic is developing quickly. Apart from these sources, there might be more relevant research, as identified in the backward research.
- Only German and English language publications were considered. Massive investments in AI research in China (see for example Roberts et al., 2021) suggest putting special emphasis on publications from the Asian region and to consider publications in Asian languages, too.
- The article is clustered in classical and new approaches. Other subdivisions might be helpful and may be considered in future research.
- The research focus with one main research question and nine inclusion keywords and three exclusion keywords can be extended in future research.
- We explicitly excluded the topic of security from the literature research process. As indicated in Schwarzl et al. (2021), a security analysis is indispensable for a comprehensive safety analysis. Thus, an enhancement of the search criteria in direction of security may complete the picture.

Data availability statement

Whilst not publicly available due to the large amount of data, the data that support the findings of this study is available on request from the corresponding author MW, moritz.waeschle@kit.edu

Author contributions

MW, FT, AB, and FP made substantial contribution to conception and design and acquisition of data. MW, FT, and AB were involved in analysis and interpretation of data and drafted the article. AA contributed during the revision. All authors contributed to the article and approved the submitted version.

Funding

A part of the work has been performed in the project AI4DI Artificial Intelligence for Digitizing Industry. This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No. 826060. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Austria, Czech Republic, Italy, Latvia, Belgium, Lithuania, France, Greece, Finland, Norway. In Austria the project was also funded by the program IKT der Zukunft of the Austrian Federal Ministry for Climate Action (BMK). The publication was partly written at Virtual Vehicle Research GmbH in Graz and partially funded within the COMET K2 Competence Centers for Excellent Technologies from the Austrian Federal Ministry for Climate Action (BMK), the Austrian Federal Ministry for Digital and Economic Affairs (BMDW), the Province of Styria and the Styrian Business Promotion Agency (SFG). The Austrian Research Promotion Agency (FFG) has been authorized for the programme management. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Acknowledgments

We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

Conflict of interest

Authors FT and FP were employed by Virtual Vehicle Research GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inform. Fusion* 76, 243–297. doi: 10.1016/j.inffus.2021.05.008

Abdi, H. (1994). A neural network primer. *J. Biol. Syst.* 2, 247–281. doi: 10.1142/S0218339094000179

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)* (Antalya), 1–6. doi: 10.1109/ICEngTechnol.2017.8308186

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., et al. (2019). "Software engineering for machine learning: a case study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in*

Practice (ICSE-SEIP) (Montreal, QC), 291–300. doi: 10.1109/ICSE-SEIP.2019.00042

Ammar, K., Pullum, L., and Taylor, B. J. (2006). "Augmentation of current verification and validation practices," in *Methods and Procedures for the Verification and Validation of Artificial Neural Networks* (Boston, MA: Springer), 13–31. doi: 10.1007/0-387-29485-6_2

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv [Preprint] arXiv:*1606.06565. doi: 10.48550/arXiv.1606.06565

Arpteg, A., Brinne, B., Crnkovic-Friis, L., and Bosch, J. (2018). "Software engineering challenges of deep learning," in *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (Prague, CR), 50–59. doi: 10.1109/SEAA.2018.00018

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature* 563, 59–64. doi: 10.1038/s41586-018-0637-6

Ayodele, T. O. (2010). Types of machine learning algorithms. *New Adv. Mach. Learn.* 3, 19–48. doi: 10.5772/9385

Bachute, M. R., and Subhedar, J. M. (2021). Autonomous driving architectures: insights of machine learning and deep learning algorithms. *Mach. Learn. Appl.* 6:100164. doi: 10.1016/j.mlwa.2021.100164

Barrett, C., Dill, D. L., Kochenderfer, M. J., and Sadigh, D. (2022). *Stanford Center for AI Safety*. Available online at: http://aisafety.stanford.edu/whitepaper.pdf (accessed August 30, 2022).

Becker, M., Lippel, J., Stuhlsatz, A., and Zielke, T. (2020). Robust dimensionality reduction for data visualization with deep neural networks. *Graph. Models* 108:101060. doi: 10.1016/j.gmod.2020.101060

Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* 1, 20–23. doi: 10.1038/s42256-018-0004-1

Berlinic, W. (2019). *What is AI Safety*? Available online at: https://wyaber.com/what-is-ai-safety/ (accessed May 1, 2022).

Bhattacharyya, S., Cofer, D., Musliner, D., Mueller, J., and Engstrom, E. (2015). "Certification considerations for adaptive systems," in *2015 International Conference on Unmanned Aircraft Systems (ICUAS)* (Denver, CO), 270–279. doi: 10.1109/ICUAS.2015.7152300

Boehm, B., Erwin, D., Ghanem, R. G., Madni, A. M., and Wheaton, M. J., editors (2018). *Disciplinary Convergence in Systems Engineering Research, 1st Edn*. Cham: Springer International Publishing.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *J. Syst. Softw.* 80, 571–583. doi: 10.1016/j.jss.2006.07.009

Burton, S. (2021). *Blog Post: Automated Driving and Safety - a Broader Perspective.* Available online at: https://www.york.ac.uk/assuring-autonomy/news/blog/safety-highly-automated-driving-complex-system/ (accessed September 17, 2022).

Burton, S., Gauerhof, L., Sethy, B. B., Habli, I., and Hawkins, R. (2019). "Confidence arguments for evidence of performance in machine learning for highly automated driving functions," in *Computer Safety, Reliability, and Security, Vol. 11699 of Lecture Notes in Computer Science*, eds A. Romanovsky, E. Troubitsyna, I. Gashi, E. Schoitsch, and F. Bitsch (Cham: Springer International Publishing), 365–377. doi: 10.1007/978-3-030-26250-1_30

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., et al. (2019). On evaluating adversarial robustness. *arXiv [Preprint] arXiv:*1902.06705. doi: 10.48550/arXiv.1902.06705

Cheng, C.-H., Diehl, F., Hamza, Y., Hinz, G., Nührenberg, G., Rickert, M., et al. (2017). Neural networks for safety-critical applications—challenges, experiments and perspectives. *arXiv:1709.00911 [cs]*.

Cheng, C.-H., Diehl, F., Hinz, G., Hamza, Y., Nührenberg, G., Rickert, M., et al. (2018). "Neural networks for safety-critical applications-challenges, experiments and perspectives," in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 1005–1006. doi: 10.23919/DATE.2018.8342158

Collopy, P., Sitterle, V., and Petrillo, J. (2020). Validation testing of autonomous learning systems. *Insight* 23, 48–51. doi: 10.1002/inst.12285

Dawson, N. (2017). *AI Safety Literature Review*. Available online at: https://bitsandatoms.co/ai-safety-literature-review/ (accessed May 1, 2022).

Department of Defense Systems (DoD) Systems (2012). *MIL-STD-882E: Department of Defense Standard Practice: System Safety*. Available online at: https://www.dau.edu/cop/armyesoh/DAU%20Sponsored%20Documents/MIL-STD-882E.pdf (accessed May 1, 2022).

Dey, A., Pal, A., and Pal, T. (2016). Interval type 2 fuzzy set in fuzzy shortest path problem. *Mathematics* 4:62. doi: 10.3390/math4040062

Dreossi, T., Ghosh, S., Sangiovanni-Vincentelli, A., and Seshia, S. A. (2017). Systematic testing of convolutional neural networks for autonomous driving. *arXiv:1708.03309*. doi: 10.48550/arXiv.1708.03309

Droegemeier, K., Kontos, C., Kratsios, M., Córdova, F. A., Walker, S., Parker, L., et al. (2019). *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*. Available online at: https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

Ebel, B. (2015). Modellierung von zielsystemen in der interdisziplinären produktentstehung, Dissertation, in: Forschungsberichte des IPEK - Institut für Produktentwicklung, ed. A. Albers, Vol.?85, Karlsruhe.

Ebert, C., and Weyrich, M. (2019). Validation of autonomous systems. *IEEE Software.* 36, 15–23. doi: 10.1109/MS.2019.2921037

Ehlers, R. (2017). "Formal verification of piece-wise linear feed-forward neural networks," in *International Symposium on Automated Technology for Verification and Analysis ATVA 2017. Lecture Notes in Computer Science(), Vol. 10482* (Cham: Springer), 269–286. doi: 10.1007/978-3-319-68167-2_19

European Commission (2020). *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics*. Technical report, European Commission.

European Commission (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment*. Publications Office, LU.

Faculty. *What is AI Safety*? Available online at: https://faculty.ai/blog/what-is-ai-safety/ (accessed July 13, 2022).

Finn, C., Levine, S., and Abbeel, P. (2016). "Guided cost learning: deep inverse optimal control *via* policy optimization," in *International Conference on Machine Learning* (New York, NY), 49–58.

Fisher, M. S. (2007). *Software Verification and Validation: An Engineering and Scientific Approach*. Springer Science & Business Media.

Folkers, A., Rick, M., and Büskens, C. (2019). "Controlling an autonomous vehicle with deep reinforcement learning," in *2019 IEEE Intelligent Vehicles Symposium (IV)* (Paris), 2025–2031. doi: 10.1109/IVS.2019.8814124

Freeman, L. (2020). Test and evaluation for artificial intelligence. *Insight* 23, 27–30. doi: 10.1002/inst.12281

Frochte, J. (2019). *Maschinelles Lernen: Grundlagen und Algorithmen in Python*. Carl Hanser Verlag GmbH Co KG. doi: 10.3139/9783446459977

Gauerhof, L., Gansch, R., Schorn, C., Schweizer, M., Heyl, A., and Rohatschek, A. (2021). *Assuring Safety of Artificial Intelligence*. Available online at: https://www.bosch.com/stories/assuring-safety-of-artificial-intelligence/ (accessed July 13, 2022).

Gausemeier, J., and Moehringer, S. (2002). Vdi 2206-a new guideline for the design of mechatronic systems. *IFAC Proc. Vol.* 35, 785–790. doi: 10.1016/S1474-6670(17)34035-1

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., et al. (2018). "Explainable AI: the new 42?," in *Machine Learning and Knowledge Extraction*, eds A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl (Cham: Springer International Publishing), 295–303. doi: 10.1007/978-3-319-99740-7_21

Goldreich, O. (2010). *P, NP, and NP-Completeness: The Basics of Computational Complexity*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511761355

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Available online at: http://www.deeplearningbook.org/ (accessed September 17, 2022).

Goodfellow, I., and Papernot, N. (2017). *The Challenge of Verification and Testing of Machine Learning*. Cleverhans-Blog. Available online at: http://www.cleverhans.io/security/privacy/ml/2017/06/14/verification.html/ (accessed September 17, 2022).

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv [Preprint] arXiv:*1412.6572. doi: 10.48550/arXiv.1412.6572

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. *Adv. Neural Inform. Process. Syst.* 29, 3909–3917.

Hagerty, A., and Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv:1907.07892 [cs]*. doi: 10.48550/arXiv.1907.07892

Han, S., Kim, J., Kim, G., Cho, J., Kim, J., and Yoo, S. (2021). "Preliminary evaluation of path-aware crossover operators for search-based test data generation for autonomous driving," in *2021 IEEE/ACM 14th*

International Workshop on Search-Based Software Testing (SBST) (Madrid), 44–47. doi: 10.1109/SBST52555.2021.00020

Hecht-Nielsen, R. (1992). "Theory of the backpropagation neural network," in *Neural Networks for Perception*, ed H. Wechsler (Elsevier), 65–93. doi: 10.1016/B978-0-12-741252-8.50010-8

Hein, M., and Andriushchenko, M. (2017). Formal guarantees on the robustness of a classifier against adversarial manipulation. *arXiv [Preprint] arXiv:*1705.08475. doi: 10.48550/arXiv.1705.08475

Hellmann, M. (2001). *Fuzzy Logic Introduction*. Rennes Cedex: Université de Rennes.

Hesenius, M., Schwenzfeier, N., Meyer, O., Koop, W., and Gruhn, V. (2019). "Towards a software engineering process for developing data-driven applications," in *2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)* (Montreal, QC), 35–41. doi: 10.1109/RAISE.2019.00014

Hirata, C., and Nadjm-Tehrani, S. (2019). "Combining GSN and STPA for safety arguments," in *Computer Safety, Reliability, and Security, Lecture Notes in Computer Science*, eds A. Romanovsky, E. Troubitsyna, I. Gashi, E. Schoitsch, and F. Bitsch (Cham: Springer International Publishing), 5–15. doi: 10.1007/978-3-030-26250-1_1

HLEG (2019). *Ethics Guidelines for Trustworthy AI*. Available online at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed May 1, 2022).

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw*. 4, 251–257. doi: 10.1016/0893-6080(91)90009-T

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., et al. (2020). A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability. *arXiv: 1812.08342*. doi: 10.1016/j.cosrev.2020.100270

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv [Preprint] arXiv:*1906.01820.

Independent High-Level Expert Group on Artificial Intelligence (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment*. Available online at: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed May 1, 2022).

International Electrotechnical Commission (2010). *IEC 61508, Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*. Geneva: International Electrotechnical Commission.

International Telecommunication Union (ITU). (2022). *Focus Group on AI for Autonomous and Assisted Driving (FG-AI4AD)*. Available online at: https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/default.aspx (accessed May 1, 2022).

ISO Central Secretary (2015). *ISO 9000: Quality Management Systems-Fundamentals and Vocabulary*. Geneva: Standard, International Organization for Standardization.

ISO Central Secretary (2018a). *ISO 21448: Road Vehicles - Functional Safety*. Geneva: International Organization for Standardization.

ISO Central Secretary (2018b). *ISO 26262-5:2018: Road Vehicles – Functional Safety – Part 5: Product Development at the Hardware Level*. Geneva: Standard, International Organization for Standardization.

ISO Central Secretary (2018c). *ISO 26262: Road Vehicles - Functional Safety*. Geneva: Standard, International Organization for Standardization.

ISO Central Secretary, International Electrotechnical Commission (2014). *ISO/IEC Guide 51:2014 Safety Aspects-Guidelines for Their Inclusion in Standards*. Geneva: Standard, International Organization for Standardization.

Juric, M., Sandic, A., and Brcic, M. (2020). *AI Safety: State of the Field Through Quantitative Lens*. Available online at: https://arxiv.org/pdf/2002.05671 (accessed May 1, 2022).

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. (2017a). "Reluplex: an efficient SMT solver for verifying deep neural networks," in *International Conference on Computer Aided Verification. CAV 2017. Lecture Notes in Computer Science(), Vol 10426.* (Springer). doi: 10.1007/978-3-319-63387-9_5

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. (2017b). Towards proving the adversarial robustness of deep neural networks. *arXiv [Preprint] arXiv:*1709.02802. doi: 10.4204/EPTCS.257.3

Kazim, E., and Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns* 2:100314. doi: 10.1016/j.patter.2021.100314

Kitchenham, B., and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University Joint Report.

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: a survey. *Int. J. Robot. Res.* 32, 1238–1274. doi: 10.1177/0278364913495721

Kong, Z., and Liu, C. (2019). *Generating Adversarial Fragments With Adversarial Networks for Physical-World Implementation*. Computing Research Repository (CoRR).

Koopman, P., Ferrell, U., Fratrik, F., and Wagner, M. (2019a). "A safety standard approach for fully autonomous vehicles," in *Computer Safety, Reliability, and Security, Vol. 11699* eds A. Romanovsky, E. Troubitsyna, I. Gashi, E. Schoitsch, and F. Bitsch (Cham: Springer International Publishing), 326–332. doi: 10.1007/978-3-030-26250-1_26

Koopman, P., Osyk, B., and Weast, J. (2019b). Autonomous vehicles meet the physical world: RSS, variability, uncertainty, and proving safety (expanded version). *arXiv: 1911.01207*. doi: 10.1007/978-3-030-26601-1_17

Koopman, P., and Wagner, M. (2016). Challenges in autonomous vehicle testing and validation. *SAE Int. J. Transp. Saf.* 4, 15–24. doi: 10.4271/2016-01-0128

Krafcik, J. (2020). *Waymo Is Opening Its Fully Driverless Service to the General Public in Phoenix*. Available online at: https://blog.waymo.com/2020/10/waymo-is-opening-its-fully-driverless.html (accessed May 1, 2022).

Krakovna, V. (2021). *AI Safety Resources*. Available online at: https://vkrakovna.wordpress.com/ai-safety-resources/ (accessed May 1, 2022).

Kurd, Z., and Kelly, T. (2003). "Establishing safety criteria for artificial neural networks," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (Springer), 163–169. doi: 10.1007/978-3-540-45224-9_24

Lee, C.-C. (1990). Fuzzy logic in control systems: fuzzy logic controller. I. *IEEE Trans. Syst. Man Cybern.* 20, 404–418. doi: 10.1109/21.52551

Li, S., Wang, W., Mo, Z., and Zhao, D. (2018). "Cluster naturalistic driving encounters using deep unsupervised learning," in *2018 IEEE Intelligent Vehicles Symposium (IV)* (Changshu), 1354–1359. doi: 10.1109/IVS.2018.8500529

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gótzsche, P. C., Ioannidis, J. P. A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 339:b2700. doi: 10.1136/bmj.b2700

Lorenzoni, G., Alencar, P., Nascimento, N., and Cowan, D. (2021). Machine learning2 model development from a software engineering perspective: a systematic literature review. *arXiv [Preprint] arXiv:*2102.07574. doi: 10.48550/arXiv.2102.07574

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv [Preprint] arXiv:*1706.06083. doi: 10.48550/arXiv.1706.06083

Mao, C., Zhong, Z., Yang, J., Vondrick, C., and Ray, B. (2019). Metric learning for adversarial robustness. *arXiv [Preprint] arXiv:*1909.00900. doi: 10.48550/arXiv.1909.00900

Masmoudi, M. S., Krichen, N., Masmoudi, M., and Derbel, N. (2016). Fuzzy logic controllers design for omnidirectional mobile robot navigation. *Appl. Soft Comput*. 49, 901–919. doi: 10.1016/j.asoc.2016.08.057

MIT (2021). *Gathering Strength, Gathering Storms*. Available online at: https://ai100.stanford.edu/sites/g/files/sbiybj18871/files/media/file/AI100Report_MT_10.pdf (accessed May 1, 2022).

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process*. 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2574–2582. doi: 10.1109/CVPR.2016.282

National Center for Statistics and Analysis. (2017). 2016 fatal motor vehicle crashes: Overview. in *Traffic Safety Facts Research Note. Report No. DOT HS 812 456).* Washington, DC: National Highway Traffic Safety Administration. Available online at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812456 (accessed May 1, 2022).

Ng, A. Y., and Russell, S. J. (2000). "Algorithms for inverse reinforcement learning," in *Proceedings of 17th International Conference on Machine Learning, Vol. 1* (Reykjavik: Morgan Kaufmann), 663–670.

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511819346

Ortega, P. A., Maini, V., and the DeepMind Safety Team. (2018). Building safe artificial intelligence: specification, robustness, and assurance.

O'Shea, K., and Nash, R. (2015). An introduction to convolutional neural networks. *arXiv [Preprint] arXiv:*1511.08458. doi: 10.48550/arXiv.1511.08458

Page, J., Bain, M., and Mukhlish, F. (2018). "The risks of low level narrow artificial intelligence," in *2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR)* (Shenyang), 1–6. doi: 10.1109/IISR.2018.8535903

Passino, K. M., Yurkovich, S., and Reinfrank, M. (1998). Fuzzy control. *IEEE Transactions on Education.* 42. doi: 10.1109/13.746327

Peri, V. M., and Simon, D. (2005). "Fuzzy logic control for an autonomous robot," in *NAFIPS 2005-2005 Annual Meeting of the North American Fuzzy Information Processing Society* (Detroit, MI), 337–342.

Psaros, A. F., Meng, X., Zou, Z., Guo, L., and Karniadakis, G. E. (2022). Uncertainty quantification in scientific machine learning: methods, metrics, and comparisons. *arXiv [Preprint] arXiv*:2201.07766. doi: 10.48550/arXiv.2201.07766

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., and Floridi, L. (2021). The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI Soc.* 36, 59–77. doi: 10.1007/s00146-020-00992-2

Ropohl, G. (2009). *Allgemeine Technologie-Eine Systemtheorie der Technik: Allgemeine Technologie-Eine Systemtheorie der Technik* 3. Überarbeitete Auflage. Karlsruhe: Universitätsverlag Karlsruhe 2009. doi: 10.26530/OAPEN_422388

Rudner, T. G., and Toner, H. (2021). Key concepts in AI safety: an overview. *Comput. Secur. J.* doi: 10.51593/20190040

Russell, S., Dewey, D., and Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Mag.* 36, 105–114. doi: 10.1609/aimag.v36i4.2577

Russell, S. J., and Norvig, P. (2022). *Artificial Intelligence: A Modern Approach, 4th Edn*. Boston, MA: Pearson.

SAE International (2021). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.* Technical Report, SAE International.

SafeTRANS, e. V. (2019a). *Safety, Security, and Certifiability of Future Man-Machine Systems.* Available online at: https://www.safetrans-de.org/de/Uploads/AK_2018_RLE_CPS/SafeTRANS_RM_SSC_FMMS_Roadmap_V2.pdf?m=1611136486 (accessed May 1, 2022).

SafeTRANS, e. V. (2019b). *Safety, Security, and Certifiability of Future Man-Machine Systems*. Available online at: https://www.safetrans-de.org/de/Uploads/AK_2018_RLE_CPS/SafeTRANS_RM_SSC_FMMS_Positionspapier_V2.pdf?m=1612514976 (accessed May 1, 2022).

Salay, R., Queiroz, R., and Czarnecki, K. (2017). An analysis of ISO 26262: using machine learning safely in automotive software. *SAE Technical Paper Series. WCX World Congress Experience.* doi: 10.4271/2018-01-1075

Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* 2, 1–21. doi: 10.1007/s42979-021-00592-x

Schroeder, D. (2010). Identifikation nichtlinearer systeme mit vorstrukturierten rekurrenten netzen. *Intell. Verfahren* 181–216. doi: 10.1007/978-3-642-11398-7_6

Schumann, H., Berres, A., Stehr, T., and Engelhardt, D. (2020). Effective selection of quality literature during a systematic literature review. *Inform. Sci. Int. J. Emerg. Transdiscipl.* 23, 77–87. doi: 10.28945/4551

Schwalbe, G., Knie, B., Sämann, T., Dobberphul, T., Gauerhof, L., Raafatnia, S., et al. (2020). "Structuring the safety argumentation for deep neural network based perception in automotive applications," in *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops. SAFECOMP 2020. Lecture Notes in Computer Science(), Vol. 12235*, eds. A. Casimiro, F. Ortmeier, E. Schoitsch, F. Bitsch, P. Ferreira (Springer, Cham), 383–394. doi: 10.1007/978-3-030-55583-2_29

Schwalbe, G., and Schels, M. (2020). "A survey on methods for the safety assurance of machine learning based systems," in *10th European Congress on Embedded Real Time Software and Systems (ERTS).* (Toulouse).

Schwarzl, C., Marko, N., Martin, H., Expósito Jiménez, V., Castella Triginer, J., Winkler, B., et al. (2021). Safety and security co-engineering for highly automated vehicles. *Elektrotech. Inform.* 7, 469–479. doi: 10.1007/s00502-021-00934-w

Serban, A., van der Blom, K., Hoos, H., and Visser, J. (2020). "Adoption and effects of software engineering best practices in machine learning," in *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (Bari), 1–12. doi: 10.1145/3382494.3410681

Seuß, D. (2021). Bridging the gap between explainable ai and uncertainty quantification to enhance trustability. *arXiv [Preprint] arXiv*:2105.11828. doi: 10.48550/arXiv.2105.11828

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2018). On a formal model of safe and scalable self-driving cars. *arXiv:1708.06374 [cs, stat*]. doi: 10.48550/arXiv.1708.06374

Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks towards data. *Science* 6, 310–316. doi: 10.33564/IJEAST.2020.v04i12.054

Singla, S., and Feizi, S. (2019). Robustness certificates against adversarial examples for RELU networks. *arXiv:1902.01235.* doi: 10.48550/arXiv.1902.01235

Skias, S. T. (2006). "Background of the verification and validation of neural networks," in *Methods and Procedures for the Verification and Validation of Artificial Neural Networks* (New York, NY: Springer), 1–12. doi: 10.1007/0-387-29485-6_1

Soares, N., and Fallenstein, B. (2017). "Agent foundations for aligning machine intelligence with human interests: a technical research agenda," in *The Technological Singularity*, eds V. Callaghan, J. Miller, R. Yampolskiy, and S. Armstrong (Berlin; Heidelberg: Springer Berlin Heidelberg), 103–125. doi: 10.1007/978-3-662-54033-6_5

Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. (2015). "Corrigibility," in *Workshop at the 29th AAAI Conference on Artificial Intelligence*. (Austin, TX).

Song, C., Cheng, H.-P., Yang, H., Li, S., Wu, C., Wu, Q., et al. (2018). "Mat: a multi-strength adversarial training method to mitigate adversarial attacks," in *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (Hong Kong), 476–481. doi: 10.1109/ISVLSI.2018.00092

Sotala, K. (2015). "Concept learning for safe autonomous AI," in *AAAI Workshop: AI and Ethics.* (Austin, TX).

Stewart, T. (2022). *Overview of Motor Vehicle Crashes in 2020*. Technical Report DOT HS 813 266, U.S. Department of Transportation's National Highway Traffic Safety Administration. Available online at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813266 (accessed May 1, 2022).

Stutz, D., Hein, M., and Schiele, B. (2019). "Disentangling adversarial robustness and generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 6976–6987. doi: 10.1109/CVPR.2019.00714

Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., and Gao, Y. (2018). "Is robustness the cost of accuracy?-a comprehensive study on the robustness of 18 deep image classification models," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 631–648.

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv [Preprint] arXiv*:1312.6199. doi: 10.48550/arXiv.1312.6199

Taylor, J., Yudkowsky, E., LaVictoire, P., and Critch, A. (2016). Alignment for advanced machine learning systems. *Ethics Artif. Intell.* 342–382. doi: 10.1093/oso/9780190905033.003.0013

Tomorrows Business GmbH (2022). *OSS.5 Europe-Operational Safe Systems.* Available online at: https://www.oss-5.com/agenda-update/ (accessed September 17, 2022).

Törngren, M. (2019). *Assurance Cases in an Era of Smart and Collaborative Cyber-Physical Systems-Pain Points and Ways Forward.* Available online at: http://www.es.mdh.se/assure2019/presentations/Martin_Torngren_ASSURE2019.pdf (accessed May 1, 2022).

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511840005

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2017). Ensemble adversarial training: attacks and defenses. *arXiv [Preprint] arXiv*:1705.07204. doi: 10.48550/arXiv.1705.07204

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2018). Robustness may be at odds with accuracy. *arXiv [Preprint] arXiv*:1805.12152. doi: 10.48550/arXiv.1805.12152

Underwriters Laboratories (2020). *UL 4600: Standard for Evaluation of Autonomous Products*. Standard for Safety. Underwriters Laboratories.

Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., and Abrahamsson, P. (2019). Ethically aligned design of autonomous systems: industry viewpoint and an empirical study. *arXiv [Preprint] arXiv*:1906.07946. doi: 10.48550/arXiv.1906.07946

VERBI Software (2021). *Maxqda 2022 Online Manual*. VERBI Software.

Walden, D. D., Roedler, G. J., Forsberg, K. J., Hamelin, R. D., and Shortell, T. M. (2015). *System Engineering Handbook, v4.0*. San Diego, CA: International Council on Systems Engineering.

Wood, M. (2019). *Safety First for Automated Driving*. Whitepaper, Aptiv Services US, LLC; AUDI AG; Bayrische Motoren Werke AG; Beijing Baidu Netcom Science Technology Co., Ltd; Continental Teves AG & Co oHG; Daimler AG; FCA US LLC; HERE Global B.V.; Infineon Technologies AG; Intel; Volkswagen AG.

Yampolskiy, R., and Fox, J. (2012). *Safety Engineering for Artificial General Intelligence*. Topoi. (Springer Nature Switzerland AG), 32, 217–226. doi: 10.1007/s11245-012-9128-9

Yang, Z., Chai, Y., Anguelov, D., Zhou, Y., Sun, P., Erhan, D., et al. (2020). Surfelgan: synthesizing realistic sensor data for autonomous driving. *CoRR, abs/2005.03844*. 11115–11124. doi: 10.1109/CVPR42600.2020.01113

Yuan, X., He, P., Zhu, Q., and Li, X. (2019). Adversarial examples: attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst*. 30, 2805–2824. doi: 10.1109/TNNLS.2018.2886017

Zadeh, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets Syst*. 11, 199–227. doi: 10.1016/S0165-0114(83)80081-5

Zadeh, L. A. (2008). Is there a need for fuzzy logic? *Inform. Sci*. 178, 2751–2779. doi: 10.1016/j.ins.2008.02.012

Zielke, T. (2020). "Is artificial intelligence ready for standardization?" in *Systems, Software and Services Process Improvement, Vol. 11699*, eds M. Yilmaz, J. Niemann, P. Clarke, R. Messnarz (Cham: Springer International Publishing), 259–274. doi: 10.1007/978-3-030-56441-4_19