Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P7168-cd

The need for systematic approaches in risk assessment of safety-critical AI-applications in machinery

Franziska Wolny[1], Silvia Vock[2], Rasmus Adler[3], Taras Holayad[4]

[1] *Federal Institute for Occupational Safety and Health (BAuA), Germany.*
E-mail: wolny.franziska@baua.bund.de
[2] *Federal Institute for Occupational Safety and Health (BAuA), Germany. E-mail: vock.silvia@baua.bund.de*
[3] *Fraunhofer Institute for Experimental Software Engineering IESE, Germany.*
E-mail: rasmus.adler@iese.fraunhofer.de
[4] *Federal Network Agency (Bundesnetzagentur), Germany. E-mail: taras.holayad@bnetza.de*

The integration of artificial intelligence (AI) into safety-critical machinery applications in industrial environments presents substantial challenges for conformity assessment and safety certification. Unlike traditional control systems, AI's data-driven nature and non-trivial behaviour complicates the assurance of compliance with established safety standards. This contribution highlights the specific challenges with respect to the new European Machinery Regulation (2023) and the AI Act (2024). We present corresponding developments in standardisation and research and discuss to what extent safety cases can underpin the safety evaluation and conformity assessment for today's applications in industry.

Keywords: risk assessment, AI, assurance case, safety case, machinery

1. Introduction

As industrial machinery increasingly incorporates artificial intelligence (AI) to enhance efficiency and autonomy, the potential risks associated with AI-driven systems in the workplace are also growing. Examples of potential safety-critical AI applications in industrial settings include safety-zone monitoring in circular saws, autonomous guided vehicles in industrial environments and autonomous systems in construction vehicles. Today, some of these applications have already been implemented as assistance functions [1, 19, 9]. In future, these applications could eventually qualify as safety functions enabling more flexible and efficient production processes. They may even facilitate production scenarios that are currently unattainable due to the limitations of traditional safety measures. However, the introduction of safety-critical systems based on machine learning (ML) in machines requires compliance with legal requirements such as the new European Machinery Regulation (MR) and the AI Act. Appropriate comprehensive risk assessment frameworks are currently being discussed in the academic and standardisation communities. While existing standards provide specific safety guidelines for conventional deterministic systems, they may not fully address the complexities introduced by AI methods. This constitutes a major challenge and is subject of this discussion paper. In section 2 we introduce challenging aspects stemming from the regulative context and the methodological point of view. Using an example, we will show how complex the interplay between the two relevant regulations (MR and AI Act) can be. In section 3 we discuss safety cases as a means of providing structured evidence of the achievement of regulatory safety objectives and give a brief overview on related standards and standardisation projects. Finally, section 4 summarises the remaining challenges and section 5 suggests possible research directions and approaches that are crucial to fill the gaps in risk assessment for AI systems.

2. Relevant legal frameworks: interplay and challenges

In this section we present the two relevant pieces of legislation for AI in machinery and discuss the role of standards in specifying the high level

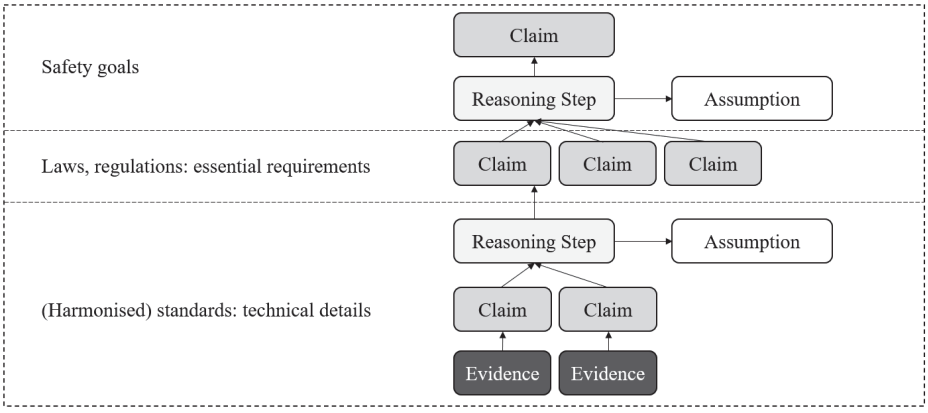

Fig. 1. Schematic depiction of the relation between safety claims and goals, regulations and the standardisation substructure in the new legislative framework. The relation between these levels can be demonstrated using a structured safety case (definitions in section 3).

requirements of these regulations.

With the MR (EU) 2023/1230 replacing the Machinery Directive 2006/42/EC and the AI Act (EU) 2024/1689 coming into force in 2024, both a revised sectoral EU product safety law and a new horizontal legal act have become crucial for the integration of safety-critical AI into machinery. Within the New Legislative Framework (NLF), the scope of these EU laws is limited to essential high-level requirements. Technical specifications are subsequently to be provided by harmonised standards. While manufacturers are not legally required to apply these standards, doing so provides a presumption of conformity with the relevant provisions of the respective legal act.

Figure 1 shows the basic idea of technical (harmonised) standards supporting safety claims of legislation. A more detailed description of the depicted schematic of a structured safety case can be found in section 3. The aforementioned legal acts define safety objectives and high-level requirements according to the risk level of a product. These requirements are addressed in technical standards in which specific measures such as procedures, tests and threshold values are defined to meet these requirements.

The EU Machinery Regulation is to be applied to the placing on the market of machines and associated products from January 20th 2027. It includes formal requirements for placing on the market as well as safety and health requirements for machinery and related products. The general principles in Annex III require the manufacturer of machinery or a related product to consider the risks that can (foreseeably) arise when placing the machinery on the market. These requirements of course also apply to AI systems embedded in machinery and related products. Additionally, in Annex I, Part A, the MR explicitly mentions the need for a special conformity assessment by notified bodies for certain machinery components containing AI. However, the MR avoids the use of the term ”artificial intelligence” and paraphrases it with ”fully or partially self-evolving behaviour using machine learning approaches”. The MR thus only explicitly mentions those systems that exhibit further development or learning after they have been placed on the market or put into operation, i.e. during use. Trained systems, that is, systems that do not continue learning, are not included in this definition [22]. Furthermore, the regulation limits the capacity of AI to systems that use machine learning approaches. Other AI methods are not explicitly considered in the MR. Annex I, Part B includes further products that are often based on AI such as protective devices designed to detect the presence of persons (No. 15) or logic units

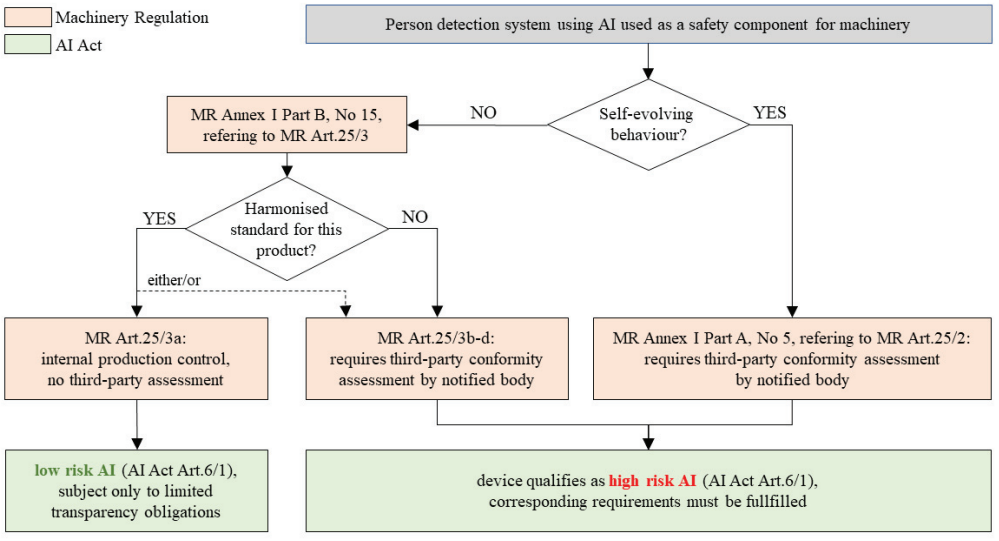

Fig. 2. Example for the application of the Machinery Regulation and AI Act to determine the necessary requirements for the placing on the market of a machinery safety component using person detection with AI.

to ensure safety functions (No. 17). For safety components and machinery that have embedded such systems, the MR may also demand a third-party conformity assessment by notified bodies. This demand constitutes one prerequisite for being classified as a high risk system by the AI Act.

The AI Act classifies AI based on risk. It lists prohibited AI practices and defines the class of 'high-risk AI systems' which are subject to special requirements. The majority of these obligations are to be met by the providers (developers) of these systems. AI systems are high-risk AI systems under Article 6(1) of the AI Act if the following two conditions are both met:

(a) the AI system is intended to be used as a safety component of a product
(b) it is required to undergo a third-party conformity assessment.

With (b), the horizontal AI Act leaves it to the sectoral (vertical) regulations, in case of machinery to the MR, whether a system with AI falls into the high-risk category.

High-risk AI systems must comply with certain requirements in the following areas (AI Act Chapter 3, Section 2):

- Risk management system
- Data and data governance
- Documentation and record-keeping
- Transparency and provision of information to deployers
- Human oversight
- Accuracy, robustness and cybersecurity

As one exemplary requirement, a risk management system is to be applied as an iterative process throughout the entire lifecycle of the AI system. Risk management shall comprise the identification and evaluation of known and foreseeable risks and the adoption of appropriate risk management measures. After that, the overall residual risk of the high-risk AI-system should be judged as acceptable.

In Figure 2, we show a schematic of selected interrelations when applying both regulations to an exemplary case, a person detection system using AI as a safety component in a machine. As mentioned above, the classification as high-risk by the AI Act depends on the necessity of a third-party conformity assessment in the MR. This is undoubtedly the case if the system exhibits self-

evolving behaviour. However, even if this does not apply, a third-party conformity assessment might still be necessary in this example if no harmonised standard for this product exists. Only if there is such a standard, the system would fall into the low risk AI category with less strict obligations. So it might depend on the progress of standardisation if a given system falls into the high-risk category with all its safety requirements or not.

As described above, the AI Act and the Machinery Regulation only formulate general safety objectives. Concrete technical details are to be defined in standards. To this end, the European Commission published a standardisation request (SR) to CEN/CENELEC in support of the AI Act and a draft of a SR for the MR. The SR in support of the AI Act includes requests for standards on risk management, quality of datasets, record keeping, transparency, human oversight, accuracy, robustness, cybersecurity, quality management and conformity assessment. The SR in support of the MR explicitly requests the development of harmonised standards. A partial overlap of the SRs can be recognised, especially for methodology to be applied covering the associated risks of emerging technology, machinery and safety components with fully or partially self-evolving behaviour and protection against corruption.

All harmonised standards must now also include an informative Annex Z which clarifies the relationship between the sections of the legal requirements and the corresponding sections of the standard. However, the Annex does not provide an unmistakable traceability between high-level regulatory requirements and the more detailed technical requirements in standards as depicted in Fig.1. A safety case can be a useful approach for arguing safety in this and many other scenarios (see definition and related work in section 3). On the one hand, it offers a structured and traceable methodology to demonstrate the safety of a product when no detailed requirements in the form of (harmonised) standards are provided. But even when technical standards are available, safety cases can be valuable if the requirements in the standards are rather high-level.

After the previous considerations, we see two major challenges which arise from the current legislative framework and state of standardisation with respect to safety:

(1) *Completeness of regulatory requirements:* It remains unclear whether the formulated procedures and requirements in MR and AI Act are sufficient and complete to claim safety. Neither the regulative texts nor its recitals justify on which basis completeness of regulatory requirements can be assumed. There is no structured argument that explains why the achievement of regulatory requirements implies safety.
(2) *Level of detail needed in the standards:* The SRs do not state the level of detail needed in the standards to fulfil the SR and the overarching safety objectives of the regulations.

3. Safety cases in research and standardisation

As mentioned in the previous section, structured safety cases can be a useful link to prove that measures proposed by standardization or chosen by AI developers lead to a sufficient level of safety for the user and fulfil the protection goals of the legislation. In this section we want to briefly introduce the concept of safety cases and give an overview of related work in research and standardisation.

Correa-Jullian et al. state that ”safety cases are a construct that serves as a framework combining claims, arguments, and the supporting evidence, justifications, and assumptions about the system’s safety. They serve multiple purposes, such as safety certification and providing structure for risk management or risk communication tasks.” [13]. Kelly et al. describe the purpose of a safety case as follows: ”A safety case should communicate a clear, comprehensive and defensible argument that a system is acceptably safe to operate in a particular context.” [20]. Graphical representations such as the Goal Structuring Notation (GSN) and Claims Arguments Evidence (CAE) are often used to visualize the safety argument in tree structure [11, 13]. In this paper, we abstract from a concrete notation like GSN or CAE and focus

on the common underlying principles and related challenges. As illustrated in Fig.1, the tree structure consists of claims. The hierarchical relation between claims means that one can conclude that a claim holds if its (lower-level) sub-claims hold. This conclusion is visualized in Fig.1 as a reasoning step. A reasoning step can come along with some assumptions meaning that the sub-claims imply the higher-level claim only if the assumptions are true. The claims at the bottom are based on evidences, that is, facts that can be proven. The top-level claim refers to the achievement of safety. The reasoning steps and the assumptions explain why and under which conditions it is possible to conclude safety from the evidences.

3.1. *Academic research*

After a thorough literature review we have the impression that there is consensus in the safety research community that safety cases are the state-of-the-art approach for assuring safety of AI systems and autonomous systems. In the following, we present relevant academic work that has led to this impression. We are aware that the ”empirical evidence for the value of safety cases is weak” [16]. However, this is true of many methods in software and dependability engineering because it is laborious and costly to conduct experiments that provide sufficient evidence to support the validity of claims.

Recent German and European research projects dealing with the assurance of AI and autonomy such as Confiance.ai [2], ExamAI [3], Absicherung KI [4] and zertifizierte KI [5] promote the usage of safety cases or directly address the challenge to come up with an appropriate safety case. The literature review by Neto et al. compiles numerous systematic approaches for safety assurance in form of safety-cases [23]. Neto et al. conclude that such approaches have the potential to accompany the development process throughout the lifecycle of safety-critical AI systems, can assist safety practitioners and go beyond the conventional V-shaped method from non-AI safety standards like IEC 61508, ISO 13849 and ISO 26262 (see Table 1 for details on standards). Concrete approaches for systematic safety-argumentation of ML components in recent years can be found in e.g. [11, 18, 14, 10, 21].

The current high abstraction level of presented safety-cases in academic literature makes the applicability for safety practitioners questionable (see also [23]). In safety standards from various application domains like ISO 26262 for automotive, safety cases are recommended and academia provides examples how such requirements can be fulfilled. This also triggered academia to provide some examples dealing with AI and autonomous systems like AI-based pedestrian detection in the context of automated driving [15]. However, use cases with a detailed safety-case application from the field of AI in machinery are not yet available even though a main recommendation of the project ExamAI [3] was to apply safety cases for AI and machinery [7].

3.2. *Standardisation Activities*

In the following, we first discuss automotive standards addressing safety cases for which we assume a potential for adoption in machinery safety. Automotive standards are particularly relevant because they tackle the challenges of complex, autonomous, and software-driven systems—issues increasingly present in modern machinery. In the second part, we mention standards specific to machinery safety and briefly discuss their relationship to the concept of safety cases.

ISO PAS 8800 considers the safety case as a core concept or the key deliverable. It describes safety-related properties of AI systems that can be used to construct a convincing safety assurance claim for the absence of unreasonable risk. This scope of ISO PAS 8800 supports the composition of an overall safety case that can be assessed by means of UL 4600. ISO PAS 8800 reflects results from the lighthouse project ”KI Absicherung” with key players from the German automotive industry.

UL 4600 intends to help ensure that an acceptably thorough consideration of safety for an autonomous product is conducted. The standard focuses on autonomous road vehicles but is based on a generalised autonomous system framework. It places emphasis on ensuring that a safety case

Table 1. Selection of relevant standards for safety and AI safety.

standard	title
ISO PAS 8800	Road vehicles – Safety and artificial intelligence
UL 4600	Standard for Safety for the Evaluation of Autonomous Products
ISO 26262	Road vehicles – Functional safety
ISO 21448	Road vehicles – Safety of the intended functionality
VDE AR 2842-61	Development and trustworthiness of autonomous/cognitive systems
ISO 13849	Safety of machinery
IEC 61508	Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems
ISO/IEC TR 5469	Artificial intelligence – Functional safety and AI systems
ISO/IEC TS 22440	Artificial intelligence – Functional safety and AI systems
ISO 12100	Safety of machinery – General principles for design – Risk assessment and risk reduction
DIN SPEC 92005	Artificial Intelligence - Uncertainty quantification in machine learning

is reasonably complete and well formed. For this purpose, compliance to standards that deal with specific aspects of safety like ISO 26262 or ISO 21448 can be incorporated into the UL 4600 safety case.

To the best of our knowledge, the only sector-overarching standard that addresses all safety aspects of AI and autonomous systems is the application rule VDE AR 2842-61 which considers three levels of abstraction. The solution level (top-level) is related to system safety. The system level relates to product safety. The technical level (lowest level) relates to functional safety. VDE AR 2842-61 considers safety cases a core concept.

In contrast to automotive and other sectors, the development of safety cases is exceptional in machinery manufacturing. They are not considered in relevant safety standards such as ISO 13849 and IEC 61508 or ISO/IEC TR 5469 and its follow-up version ISO/IEC TS 22440 (both are driven by the maintenance group of IEC 61508).

ISO 12100 is the key standard for providing a comprehensive framework for systematically identifying hazards, assessing risks, and implementing effective measures to ensure machinery safety throughout its lifecycle. It does not address safety case development but its principles provide foundational elements and process steps that could be incorporated into a safety case.

In conclusion, safety cases play a pivotal role in sectors like automotive, where standards such as UL 4600 and ISO PAS 8800 emphasize their importance for ensuring the safety of complex and autonomous systems. While machinery safety standards like ISO 12100 provide essential principles for risk assessment and mitigation, incorporating safety cases into this domain could enhance the rigour and transparency of safety justifications, particularly for addressing AI and autonomy.

4. Research challenge

According to the conclusions of sections 2 and 3, one of the challenges in achieving and arguing safety of machines embedding AI is the unstructured and to date partly incomplete collection of requirements in regulations, standardisation requests and standards. The interplay between different regulations, i.e. MR and AI Act, and the interplay between different standards complicates the situation even more. Researchers propose safety cases for dealing with this issue and for assuring safety of AI-based autonomous systems (see section 3.1). This state-of-art approach is already reflected in some innovative safety standards (see section 3.2). They include for instance guidance on the topics that a safety case should address, advice for dealing with these topics and pitfalls that shall be avoided. However, compliance to this guidance can hardly guarantee that the resulting safety argument is sufficiently strong

and will be accepted by notified bodies. In the following, we first explain the underlying research challenges to address this issue and then make a proposal to address them.

A first challenge is the structure of safety arguments. Researchers have proposed different structures and there is currently no consensus which structure should be used under which conditions. For instance, a popular approach is given by AMLAS (Assurance of Machine Learning for use in Autonomous Systems) [6]. AMLAS incorporates a set of safety case patterns and a process for systematically integrating safety assurance into the development of ML components. However, Klaes et al. propose a different approach due to problems identified in applying AMLAS [21]. One problem was that claims with the term ”sufficient” are decomposed in such a way that the ”sufficiency” in subclaims is interdependent, so that the tree structure does not follow a divide and conquer approach. For example, a claim about a sufficient level of safety is decomposed over many steps into claims about a sufficient level of training data quality and a sufficiently correct trained model.

A second challenge concerns the lack of knowledge about the technology and the context in which it is applied. The application of machine learning leads for instance to several kinds of uncertainties described in DIN SPEC 92005. A typical uncertainty regarding the application context is, for example, the likelihood of safety-critical situations that the system has to handle. An approach to deal with this challenge is to identify this lack of knowledge by assessing the reasoning steps of a safety case and to collect related field data that addresses the assumptions in the reasoning. This can strengthen the argument. The growing argument strength can then be used to steer a stepwise market introduction. For instance, market introduction may begin with a few demonstrators with a human safety supervisor as we see it currently in the case of robo-taxies. If the safety argument becomes stronger due to a positive evaluation of assumptions and additional evidences for claims, then we can argue that more products with less human supervision are acceptable. Contributions to this ramp-up approach are presented in [8, 12, 18, 17]. The extent to which these approaches can be combined into an overall solution applicable to the safety of autonomous machines is still an open research question.

5. Proposal: Community driven exemplary safety case development

Our proposal to deal with the challenges presented in the previous section is to develop a reference example and establish an open community of practice around this example. Interested individuals, companies and institutions can join and raise concerns about the arguments in the example. This community is invited to introduce names for types of arguments, collect counterarguments and reactions on these counterarguments. The safety community already uses terms like ”proven-in-use argument” and there are many discussions around this kind of argument such as [24]. Our idea is to initiate similar discussions in an open community of practice for safety assurance of AI systems and autonomous systems. Researchers with solution approaches such as AI robustness analyses can use the example to explain at which points their solution is needed to strengthen a given safety argument. An additional goal is to provide different variants of the example and to let the community comment on the preferred representation of the same argument (different structure but same top-level conclusion and evidences) or the strength of different arguments.

Summarizing, as concrete next steps we first plan to find a promising use case from the autonomous machinery domain. We encourage our valued readers to contact us with information on suitable applications. From this, we aim to create the example and its variants including names for different types of arguments and different structuring approaches. Different approaches will be developed to gradually increase the strength of arguments by monitoring the fulfilment of claims and assumptions using field data. Different approaches will be developed to formalise uncertainty and strength of arguments.

Acknowledgement

The authors would like to thank Thomas Mössner for his valuable input.

References

1. https://www.altendorfgroup.com/en/world-first-hand-guard-protects-the-most-important-thing-in-trade-the-hands/. [Online; accessed 08-January-2025].
2. https://www.confiance.ai/. [Online; accessed 08-January-2025].
3. https://testing-ai.gi.de/english. [Online; accessed 08-January-2025].
4. https://www.ki-absicherung-projekt.de/en/. [Online; accessed 08-January-2025].
5. https://www.zertifizierte-ki.de/. [Online; accessed 08-January-2025].
6. https://www.assuringautonomy.com/amlas. [Online; accessed 08-January-2025].
7. R. Adler and M. Kläs. Assurance cases as foundation stone for auditing AI-enabled and autonomous systems. In *HCI International 2022 – Late Breaking Papers: HCI for Today's Community and Economy. Lecture Notes in Computer Science*, volume 13520. Springer, Cham., 2022.
8. P. Bishop, A. Povyakalo, and L. Strigini. Bootstrapping confidence in future safety from past safe operation. In *IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*, pages 97–108, 2022.
9. P. Boden, S. Rank, and T. Schmidt. Scheduling automated guided vehicles considering transport load transfers. *Logistics Research*, 16(1), 2023.
10. M. Borg, J. Henriksson, K. Socha, O. Lennartsson, E. Sonnsjö Lönegren, T. Bui, P. Tomaszewski, S. R. Sathyamoorthy, S. Brink, and M. Helali Moghadam. Ergo, SMIRK is safe: a safety case for a machine learning component in a pedestrian automatic emergency brake system. *Software quality journal*, 31(2):335–403, 2023.
11. S. Burton and B. Herd. Addressing uncertainty in the safety assurance of machine-learning. *Frontiers in Computer Science*, 5, 2023.
12. A. Cavalcanti and J. Baxter. Confidence in Assurance 2.0. In *The Practice of Formal Methods. Lecture Notes in Computer Science*, volume 14780. Springer, 2024.
13. C. Correa-Jullian, J. Grimstad, S. A. Dugan, M. A. Ramos, C. A. Thieme, A. Morozov, I. B. Utne, and A. Mosleh. *Proceedings of the 4th International Workshop on Autonomous Systems Safety*. The B. John Garrick Institute for the Risk Sciences, 2023.
14. J. P. C. de Araujo, B. V. Balu, E. Reichmann, J. Kelly, S. Kugele, N. Mata, and L. Grunske. Applying concept-based models for enhanced safety argumentation. In *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*, pages 272–283. IEEE, 2024.
15. L. Gauerhof, P. Munk, and S. Burton. Structuring validation targets of a machine learning function applied to automated driving. In *Computer Safety, Reliability, and Security (SAFECOMP 2018) Lecture Notes in Computer Science.*, volume 11093, pages 45–58, 2024.
16. I. Habli, R. Alexander, and R. D. Hawkins. Safety cases: An impending crisis? In *Safety-Critical Systems Symposium (SSS'21)*, 2021.
17. R. Hawkins and P. Ryan Conmy. Identifying run-time monitoring requirements for autonomous systems through the analysis of safety arguments. In *Computer Safety, Reliability, and Security: 42nd International Conference, SAFECOMP 2023*. Springer, 2023.
18. B. Herd, J.-V. Zacchi, and S. Burton. A deductive approach to safety assurance: Formalising safety contracts with subjective logic. In *International Conference on Computer Safety, Reliability, and Security*, pages 213–226. Springer, 2024.
19. I. Jang, J. Kim, D. Lee, C. Kim, C. Oh, Y. Kim, S. Woo, H. Sung, and H. J. Kim. Towards fully integrated autonomous excavation: Autonomous excavator for precise earth cutting and onboard landscape inspection. *IEEE Robotics Automation Magazine*, pages 2–16, 2024.
20. T. Kelly and R. Weaver. The goal structuring notation–a safety argument notation. In *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*, volume 6. Citeseer Princeton, NJ, 2004.
21. M. Kläs, R. Adler, L. Jöckel, J. Groß, and J. Reich. Using complementary risk acceptance criteria to structure assurance cases for safety-critical AI components. In *AISafety@ IJCAI*, 2021.
22. T. Mössner and M. Kittelmann. Maschinen mit Kuenstlicher Intelligenz - Anforderungen an Hersteller nach der neuen EU-Maschinenverordnung und Bezüge zur KI-Verordnung. *ARP (preprint)*, 2025.
23. A. V. S. Neto, J. B. Camargo, J. R. Almeida, and P. S. Cugnasca. Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work. *IEEE Access*, 10:130733–130770, 2022.
24. H. Schaebe and J. Braband. Basic requirements for proven-in-use arguments. *preprint*, 2015.