Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
© 2026 ESREL 2026 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p28974-cd

Arguing Safety for AI-driven Systems

Mark Locherer
DHBW Ravensburg and University of Siegen, Germany. E-mail: locherer@dhbw-ravensburg.de

Michael Kordovan
ifm electronic gmbh, Germany.

Bernd Buxbaum
University of Siegen and ifm electronic gmbh, Germany.

Thorsten Kever
DHBW Ravensburg, Germany.

The first international AI safety standards, namely ISO/IEC TR 5469 and ISO/PAS 8800, have been released. The usage of AI in safety-critical settings offers new dimensions, including autonomous mobile robots, self-driving cars, perception systems etc. Traditional safety engineering primarily employs the V-model process, as a systematic means for translating requirements from specification to system hardware and software architecture and measuring their correct implementation during verification and validation. In contrast, AI-driven systems are predominantly data-driven and follow an iterative lifecycle. Within this lifecycle, the core objective is to compile an assurance argument that demonstrates the system's capability to implement necessary risk mitigations and address the inherent uncertainties of AI-based systems. The efficient assurance argument has been proposed as a means to characterize the key facets, necessary for system assessment, deployment readiness and certifiability. This study firstly, analyzes the first two international standards to determine their alignment with these characteristics and secondly identifies open questions, particularly the challenge of transitioning safety requirements assigned to AI-based safety-critical systems into quantifiable metrics.

Keywords: Assurance argument, functional safety, artificial intelligence, machine learning, standardization, certification, safety-critical system.

1. Introduction

The first standards in addressing the usage of artificial intelligence (AI) to achieve functional safety (FuSa) have been released. Likewise, European Union regulation is considering the usage of AI generally in the horizontal AI Act (AIA) and more specifically in the machinery sector in the Machinery Regulation (MR) European Parliament, Council of the European Union (2024, 2023). The promise is to make use of AI technology in safety-related context, such as autonomous guided vehicles, mobile robots, safe human detection and many more. The common denominator of safety-critical systems (SCSs) is the necessity to using dedicated safety engineering standards, such as DIN (2023) for driverless industrial trucks, and obligatory third-party system assessment Burden et al. (2024).

On an international level, ISO/IEC TR 5469 ISO/IEC (2024) and ISO/PAS 8800 ISO (2024) have been published, addressing the usage of AI technology in general and for vehicles, respectively. Traditional safety engineering is based on the structured V-model process IEC (2010); ISO (2018); CEN (2023), ensuring a system implementation fulfilling requirements, dedicated usage. Further, it forms the basis for traceable third-party assessment. Conversely, AI based systems are iteratively developed due to their data-driven nature, i.e., a model is repeatedly refined using (updated) data, model training and testing Perez-Cerrolaza et al. (2024).

AI based SCSs, falling under Annex I(A) in the MR and under Annex I(A) in the AIA require

mandatory notified body assessment Locherer et al. (2026). However, traceability is not given when developing AI based systems, i.e., complex black box model structures impede transparency and understanding Rudin (2019); ISO/IEC (2024). An a priori specification is also not present due to unique tasks Spanfelner et al. (2012); Koopman and Wagner (2016); ISO/IEC (2024). Moreover, data is used as information source, which might not contain the relevant details. In this context, uncertainty originates from the open environment, the imprecise observations, and the model based system Kläs and Vollmer (2018); Burton and Herd (2023).

To address these challenges, assurance arguments (AAs) have been proposed to demonstrate that a system is acceptably safe Picardi et al. (2019); Ward and Habli (2020); Burton and Herd (2023). An AA is thus, a systematic composition of individual claims, arguments and evidences detailing why such a system is safe ACWG (2021). While the construction of a compelling AA is challenging Shankar et al. (2022); Annable et al. (2026), minimizing the "assurance uncertainty" is even more complex, i.e., an AA that properly mirrors the actual system uncertainty, as identified in Burton and Herd (2023). This study focuses on the "efficient AA," which supports smooth third-party assessment during certification, as detailed in Locherer et al. (2026).

In this paper, we conduct a variance comparison between implemented strategies in international standardization meeting the theoretical properties of the efficient AA. This comparison is realized by firstly detailing the key characteristics of the efficient AA, secondly, mapping relevant strategies outlined in ISO/IEC TR 5469 and ISO/PAS 8800 to this concept to identify fulfilling and lacking properties.

With our contribution, we firstly, determine the state-of-the-art of current safe AI standardization on an international level in assisting third-party assessment, secondly, we identify limitations, potentially leading to further research, and thirdly, we describe how both standards documents can be used jointly.

We use the following conventions in this manuscript: ISO/IEC TR 5469 is abbreviated with technical report (TR) and ISO/PAS 8800 with international standard (IS).

Our manuscript is structured as follows: Section 2 introduces frequently used terms, as well as the concept and the motivation behind the efficient AA. Section 3 outlines our approach and research questions (RQs). Section 4 illustrates the mapping between the TR and the IS to the properties of the efficient AA. The RQs are answered in Section 5. In Section 6 we provide a short conclusion on our study.

2. Background

This section introduces terms, system considerations and the efficient AA. Some of the terms and concepts are unique in the TR and IS.

2.1. *Application and usage level*

The TR defines the application and usage level (AUL) as a mechanism to distinguish between different usage contexts and autonomy levels of AI technology. Level A relates to using AI inside a safety related function and level B is assigned to using AI technology as tool during development (cf. ISO/IEC, 2024, Cl. 6).

2.2. *AI technology class*

Additionally, the TR introduces a classification scheme for AI-based SCSs into Class I, Class II and Class III. A Class I systems shows properties that can be addressed using existing conventional standards. Class II systems can partially be developed based on traditional safety standards, however alternative FuSa risk assessment methods are used to demonstrate compliance (cf. ISO/IEC, 2024). Class III systems cannot be implemented.

2.3. *Safety-related properties*

Safety-related properties are a concept in both standards documents, which incorporates the challenges of AI-based systems used to implement risk mitigation in a safety context. Due to the properties of AI-based systems, some limitations will likely remain in an AI-based SCS even after rigorous engineering. Hence, the idea is to com-

pensate these remaining limitations with safety-related properties. Examples are: Reliability, robustness, resilience and others. While the IS uses the term "safety-related property," the TR uses the notion "desirable property" ISO (2024); ISO/IEC (2024).

2.4. *System view and boundaries*

An AI-based SCS usually consists of hardware (HW), software (SW) and AI-based SW parts ISO/IEC (2024); ISO (2024). This is called the "encompassing system" in ISO (2024), to which we refer to as broader system or overall system. In this manuscript, we assume that a compelling AA can be developed for the conventional HW and SW parts. Furthermore, interfaces between traditional SW and AI-based SW are clearly specifiable. Additionally, we assume that the inference HW (cf. "AI technology elements," e.g., a graphics processing unit etc. in (ISO/IEC, 2024, Cl. 7)) for the safety-related AI SW is acceptably safe. During design the safety-requirements are distributed among the above mentioned parts, hence, a subset of the safety-requirements is handed over to the AI part of the SCS, for which an efficient AA shall be developed. Therefore, the AA is limited to the AI-based SW.

2.5. *Efficient assurance argument*

The "efficient AA" characterizes an AA that can be used by third party assessors, such as notified bodies. Its characterization is presented in Locherer et al. (2026), which is based on ACWG (2021); Shankar et al. (2022); Burton (2024), with the following features:

(a) The usage of a robust assurance driven development (ADD) process, as well as, qualified and recognized tools along the AI system lifecycle, i.e., "assurance driven development and reusable artifacts."
(b) The translation from descriptive safety requirements into assessable and measurable AI model properties must be given, i.e., "precise claims."
(c) Transparency in the AI technology, task and assumptions must also be given, i.e., denoted as "validatable models and assumptions."
(d) A sound decomposition structure in the AA is present, enabling simple assessment, i.e., "rigorous chain of reasoning and evidence."

3. Methodology

We employ a derived approach, similar to the "mapping study" in Petersen et al. (2015). However, instead of a conventional literature review, we base our mapping solely on ISO/IEC TR 5469 and ISO/PAS 8800 and the definition of the efficient AA. In existing literature Shankar et al. (2022); Burton and Herd (2023); Burton (2024), the efficient AA has been described mainly as a means for effective and realistic third-party SW assessment (cf. Section 2.5). This is required for SCSs and specifically for AI-based SCSs due to their inherent complexities. We map the relevant provisions detailed in ISO/IEC TR 5469 and ISO/PAS 8800 to the key properties of the efficient AA. Besides providing an overview on the different approaches detailed in the TR and the IS, we address the following RQs: RQ 1: What strategies are outlined in the TR and the IS to meet the criteria of the efficient AA? RQ 2: To what extent do both standards documents fulfill the definition of the efficient AA? RQ 3: What are the potential weaknesses and limitations in both documents with respect to the efficient AA? RQ 4: How can the TR and IS benefit from each other?

4. Mapping results

In this section the mapping is conducted. Besides providing on overview and more details on the efficient AA, we address RQ 1.

4.1. *Assurance driven development and reusable artifacts*

A SW lifecycle approach focused on creating SW and associated evidences for the AA compilation. Approved, trusted, and well-understood tools and methods applied during development.

ISO/IEC TR 5469 The TR does not ship with a predefined AI lifecycle process. Although based on IEC 61508-3, it has to be selected by the engineering team. Further, the process of the "three-stage realization principle" to incorporate

for data must be connected to the chosen lifecycle, as noted in Locherer et al. (2025). ADD is not given, and thus, must be defined based on existing international standards (ISO/IEC, 2024, Cls. 11.1–11.4, Annex D) and the three-stage realization principle (ISO/IEC, 2024, Cl. 7, Annex B). Considerations on reusable artifacts given as "AI technology elements" are presented in (ISO/IEC, 2024, Cl. 7). The definition of the AUL comprises tooling (levels B1 and B2) for which tool qualification might be necessary (ISO/IEC, 2024, Cl. 6). Special remarks on achieving compliance with AI-based tooling can be found in Slotosch (2025).

ISO/PAS 8800 The IS provides an ADD lifecycle for AI systems from requirements allocation, development to operations, as well as dataset curation (ISO, 2024, Cls. 7, 11.4). In both cases, the iterative lifecycle ensures that development and assurance are closely linked. This implies that the prerequisite for obtaining a sound AA is present. Moreover, the standard clearly states that requirement artifacts from "past products" can be recycled (ISO, 2024, Cl. 9.4), which shows that the idea of "reusable tools/artifacts" is implemented. Additionally, requirements are defined to prevent errors and confidently use external tooling and methods along the lifecycle (ISO, 2024, Cl. 15). An overview of general architectural considerations and development measures is also provided (ISO, 2024, Annex G).

4.2. *Precise claims*

Safety requirements translatable into measurable properties of the AI model.

ISO/IEC TR 5469 The requirements elicitation is given via the three-stage realization principle (ISO/IEC, 2024, Cl. 7, Annex B). This method is used to convert safety requirements into safety-related properties (ISO/IEC, 2024, Cl. 8) and to find an associated argumentation, in which AI metrics (ISO/IEC, 2024, Cl. 9.3.4) are used. The TR explicitly states that multiple metrics and field monitoring might be required, which perfectly aligns with the idea of safety performance indicators (SPIs) in UL (2023) and safety-related key performance indicators in ISO (2024). Additionally, architectural considerations are outlined under (ISO/IEC, 2024, Cl. 10) to obtain robustness and safety. However, as noted the challenge lies in "[...] defining and guaranteeing their reliability properties and failure behaviours[sic]" (ISO/IEC, 2024, Cl. 10.1).

ISO/PAS 8800 The IS focuses on how safety requirements are formulated and refined during development. Compared to the TR a whole clause is dedicated to this topic and many insights are provided (ISO, 2024, Cl. 9). An overview on safety-related properties, required for safety applications is detailed in (ISO, 2024, Annex D). Further guidance is delivered on measures, e.g., system architecture considerations, redundancy, data aspects, monitoring, among others (ISO, 2024, Cl. 14, Annex G). A note on well-known machine learning (ML) performance metrics is also given in (ISO, 2024, Annex H). The responsibility for defining a complete set of AI safety requirements and specifying them in a quantifiable manner, such as measuring an AI model's properties, lies with the engineering team.

4.3. *Validatable models and assumptions*

Explainable models and a complete understanding of the task and the environment is probably the most challenging when it comes to AI and specifically ML. While explainable models based on AI technology might never be available Rudin (2019), the engineering team must ensure that the assumptions such as task and environment are properly understood and AI system limitations are identified. In this context, ADD helps to scan the environment to precisely define the operational design domain (ODD), to obtain deeper and clearer understanding of possible scenarios imposing challenges to the model.

ISO/IEC TR 5469 The TR discusses properties and "risk factors" of AI systems in (ISO/IEC, 2024, Cl. 8), which is based on existing literature. The TR makes a distinction between training algorithm, the (untrained or empty) model and

its parameters (trained model). In some cases the model's parameters can be engineered manually, in which domain knowledge is used to tweak parameters, leading to a Class I SCS. Conversely, when an ML algorithm is used to derive the model parameters based on a dataset, this likely results in either Class II or Class III systems, necessitating alternative FuSa risk assessment methods. Further considerations refer to system autonomy, as well as, transparency and explainability as safety-related properties.

The TR underscores the importance of data when it comes to verification and validation (V & V). Hence, generating a dataset based on hazard analysis and risk assessment (HARA) should result in an "as precisely as possible" defined ODD, and establishing metrics for their verification (cf. SPIs). Likewise, reverse engineering the model to make it understandable is suggested, potentially resulting in a Class I system. Moreover, different testing procedures are proposed such as virtual and physical testing to identify the realized risk mitigation (ISO/IEC, 2024, Cl. 9.3).

ISO/PAS 8800 The IS provides a sound ADD process covering the complete lifecycle, including post deployment. It aims at generating evidences that the AI-based SCS is and remains safe. V & V is continuously conducted to ensure that safety requirements are fulfilled appropriately (ISO, 2024, Cl. 12). The iterative nature allows to "[...] identify, analyse, reduce and mitigate functional insufficiencies in the trained model [...]" (ISO, 2024, Cl. 7.5). Likewise, the definition and selection of data is emphasized in this process (ISO, 2024, Cl. 7). Entangled with the ADD is the creation of the AA, as outlined in (ISO, 2024, Cl. 8). The IS provides decomposition of the top level claim into sub claims, strategies and evidences using goal structuring notation in (ISO, 2024, Annex B).

The key objective is to demonstrate that safety requirements are fulfilled and to assess whether the risk detailed in the AA resembles the actual risk of the AI-based SCS. This is also achieved using a multiple testing approaches as outlined under (ISO, 2024, Cl. 12), including statistical testing, random testing and virtually generated test cases, among others. Moreover, the process of obtaining safety requirements requires a well understood ODD and task. Within the ADD process, refinement of requirements and ODD takes place, based on newly gained insights. Identified restrictions are renegotiated within the broader system context (ISO, 2024, Cl. 9).

4.4. *Rigorous chain of reasoning and evidence*

Assessability of the evidence and chain of arguments must be given. Evidence must be comprehensible and clear as noted for Item b.

ISO/IEC TR 5469 The TR characterizes AI-based systems using an AUL (ISO/IEC, 2024, Cl. 6), to incorporate not only for AI technology in safety-related functionality implementing risk mitigation, but also during development, e.g., for automated code generation, etc. As argued in Locherer et al. (2025), this high-level distinction offers a simple means to chose the required rigor in safety-assurance. The considerations outlined under (ISO/IEC, 2024, Cl. 9)—V & V—such as HARA and related data; different levels of testing, such as virtual and physical testing; field monitoring, etc. might support evidence generation and safety case argumentation. Very little is mentioned on how to reason for AA completeness using the "three-stage realization principle."

ISO/PAS 8800 As noted, the engineering and maintainability of the automated driving system is based on ADD. In this context, the process depicted under (ISO, 2024, Cl. 7) ensures that violations of safety requirements due to functional insufficiencies are either omitted or mitigated over the complete AI lifecycle (ISO, 2024, Cl. 8, Annex B). Likewise, evidence is generated to demonstrate the validity of this reasoning, which is based on the AI safety requirements (ISO, 2024, Cl. 9), operational measures (e.g., monitoring) (ISO, 2024, Cl. 14, Annex B) and their continuous V & V (ISO, 2024, Cls. 12, 13) using SPIs.

5. Discussion

In this section we provide our evaluation and interpretation on the presented RQs.

5.1. *Fulfillment and limitations in meeting the efficient AA*

This section details RQ 2 and RQ 3. Items a to d relate to the definition of the efficient AA under Section 2.5.

Item a: The TR does not provide an ADD process. A process has to be defined based on existing international standards and must be enriched by the guidance given in the TR. Some considerations on reusable artifacts are provided. Conversely, the IS provides a well-defined ADD, addressing ML-based SCSs. Requirements on reusable artifacts are given.

Item b: Challenges are stated in the TR, however, mitigation approaches seem ambiguous. Very little guidance is given on how to define safety-requirements, select relevant safety-related properties and which metrics have to be used for their confirmation. Further, how to measure the required amount of risk reduction based on individual or combined evidences? The IS makes provisions on the ADD, requirements elicitation, metrics etc. However, what is a consistent and complete set of safety requirements? How are multiple measures combined to substantiate a claim? While the IS provides more details and guidance, the issue in defining precise claims is assumed to be known. This however, is usually not the case (Werling et al., 2025, Sec. 1.2). Our rating is partially addressed in both, while the IS is more comprehensive.

Item c: The TR specifies problems and possible solutions by making models understandable, potentially realizable by simple models. Nevertheless, safety assurance for perception systems is given via the three-stage realization principle. In this regard, our considerations under the previous item are relevant. The challenge is to define a complete set of safety requirements, deriving safety-related properties and consistently measuring their correctness. Are the right metrics selected? Do they measure the appropriate properties, etc.? When using testing for evidence generation, what is sufficient testing? What has to be tested? Since, ADD is not given in the TR, convincing a notified body is challenging not adopting an ADD process from other standards. The IS also addresses the problem by generating evidences during the lifecycle. The key point is mentioned, which is to detail required risk mitigation and likewise correct reflection in the AA. The iterative nature of refinement and safety analysis is aimed to result in a solid solution. Based on our findings, the IS is more comprehensive compared to the TR. It addresses relevant aspects and provides a reasonable way to achieve the unachievable, by firstly, doing as much as possible and providing measures to ensure system correctness. As this problem is still not resolved, we consider it partially fulfilled.

Item d: The AI technology class conversion mechanism in the TR is incompletely described, specifically when converting a system assigned Class III, i.e., prohibited usage in a safety function, into Class II, i.e., allowed for usage. Clear requirements on what is enough evidence are missing. Hence, argument acceptance is left to the judgment of the engineering team. In this regard, a notified body might conclude differently (cf. Locherer et al. (2025)). Although the three-stage realization principle provides a means to determine evidences for selected safety-related properties, it is assumed that the proper decomposition structure, demonstrating freedom of unreasonable risk, is self-evident, which is not the case. From an assessor's perspective, is it clear which technology class should be selected? How to ensure that measures do not lack the same blind spots? Is testing economically practical? How to demonstrate that assurance uncertainty is mitigated? Are the implemented monitors providing correct and ongoing evidences? The IS systematically approaches the considerations in Item d. However, arguing for evidence completeness is challenging and makes third party assessment difficult. Therefore, more guidance, e.g., in the shape of a systematic approach to uncertainty quantification, and examples are required to prevent the development of AAs for which "[n]o evaluator would have the resources to draw out all of the flaws in such a complex assurance case[,]" as indicated in Shankar et al. (2022).

5.2. *Interim summary and recommendations on mutual benefits*

This section provides an interim summary, on which we provide a recommendation on RQ 4.

In summary, the TR outlines important aspects on AI-based FuSa, but the ADD, as a precondition for obtaining an efficient AA is missing. Hence, it could benefit from adapting or referencing to the IS. Conversely, the IS fulfills this requisite. When it comes to safety assurance, both standards expect, that a complete set of safety requirements for the AI SCS has been derived, all relevant safety-related properties have been identified and each property has been connected to a set of metrics and measures, capable of reliably, accurately and completely verifying these properties. This however, is probably the most challenging part and not obvious. Some of the safety-related properties might be unachievable or if they are achievable, providing evidence on their fulfillment is challenging. While the scope of the IS is limited to constructing an AI-based safety function with tolerable risk level, the TR can be used for considerations on tooling utilizing the AUL definition.

6. Conclusion

This study provides an overview on methods detailed in ISO/IEC TR 5469 (TR) and ISO/PAS 8800 (IS) in meeting the concept of the efficient AA. This overview is detailed based on a mapping of the methods in both standards documents to the related properties of the argument. In a nutshell, neither of the standards fulfills the notion of the efficient AA. However, the IS provides a more compelling set of methods, specifically its well-documented ADD. Conversely, the TR shows significant gaps. Hence, the condition for obtaining a sound AA can be seen as fulfilled in the IS, while the other positions are challenging to achieve using the prescribed strategies. The TR, however, meets the criteria insufficiently. Nevertheless, we believe that both standards can benefit from each other. This is specifically, given in using the TR for AI-based tooling and utilizing the ADD from the IS in the TR. The main challenge, namely providing evidence that the system is acceptably safe in a traceable manner remains.

Acknowledgement

This research did not receive any funding.

References

ACWG (2021). SCSC-141C: Goal Structuring Notation Community Standard.

Annable, N., M. Lawford, R. F. Paige, and A. Wassyng (2026). Principled Safety Assurance Arguments. In *Computer Safety, Reliability, and Security*, Cham, pp. 18–32. Springer Nature Switzerland.

Burden, H., S. Stenberg, and K. Flink (2024). When AI meets machinery – the role of the notified body. Technical Report 2024:56.

Burton, S. (2024). Safety under uncertainty: Automotive standards for AI safety and research perspectives. Presented at ”Incontri del Giovedì” by IEIIT CNR.

Burton, S. and B. Herd (2023). Addressing uncertainty in the safety assurance of machine-learning. *Frontiers in Computer Science 5*, 1132580.

CEN (2023). EN ISO 13849-1: Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design.

DIN (2023). DIN EN ISO 3691-4: Industrial trucks Safety requirements and verification Part 4: Driverless industrial trucks and their systems (ISO 3691-4:2023); German version EN ISO 3691-4:2023.

European Parliament, Council of the European Union (2023). Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery.

European Parliament, Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence.

IEC (2010). IEC 61508-1: Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 1: General requirements.

ISO (2018). ISO 26262-1: Road vehicles – Functional safety – Part 1: Vocabulary.

ISO (2024). ISO/PAS 8800: Road vehicles – Safety and artificial intelligence.

ISO/IEC (2024). ISO/IEC TR 5469: Artificial Intelligence – Functional safety and AI systems.

Kläs, M. and A. M. Vollmer (2018). Uncertainty in Machine Learning Applications: A Practice-Driven Classification of Uncertainty. In *Computer Safety, Reliability, and Security*, Cham, pp. 431–438. Springer International Publishing.

Koopman, P. and M. Wagner (2016). Challenges in Autonomous Vehicle Testing and Validation. *SAE International Journal of Transportation Safety 4*(1), 15–24.

Locherer, M., M. Kordovan, B. Buxbaum, and T. Kever (2025). No safe AI standardization in sight at present: An investigation into ISO/IEC TR 5469. In *2025 9th International Conference on System Reliability and Safety (ICSRS)*, Turin, Italy. IEEE. in press.

Locherer, M., M. Kordovan, B. Buxbaum, and T. Kever (2026, February). The missing link to safe AI homologation. In *IASEAI'26*, Paris, France. International Association for Safe and Ethical AI. in press.

Perez-Cerrolaza, J., J. Abella, M. Borg, C. Donzella, J. Cerquides, F. J. Cazorla, C. Englund, M. Tauber, G. Nikolakopoulos, and J. L. Flores (2024). Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. *ACM Comput. Surv. 56*(7), 176:1–176:40.

Petersen, K., S. Vakkalanka, and L. Kuzniarz (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology 64*, 1–18.

Picardi, C., R. Hawkins, C. Paterson, and I. Habli (2019). A Pattern for Arguing the Assurance of Machine Learning in Medical Diagnosis Systems. pp. 165–179.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Comment: Author's pre-publication version of a 2019 Nature Machine Intelligence article. Shorter Version was published in NIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning. Expands also on NSF Statistics at a Crossroads Webinar.

Shankar, N., D. Bhatt, M. Ernst, M. Kim, S. Varadarajan, S. Millstein, J. Navas, J. Biatek, H. Sanchez, A. Murugesan, and H. Ren (2022). DesCert: Design for Certification. Comment: 142 pages, 63 figures.

Slotosch, O. (2025, November). Safety Compliance and Confidence into AI Agents. In *2025 9th International Conference on System Reliability and Safety (ICSRS)*, Turin, Italy. IEEE. in press.

Spanfelner, B., Detlev Richter, S. Ebel, U. Wilhelm, and C. Patz (2012). Challenges in applying the ISO 26262 for driver assistance systems.

UL (2023). ANSI/UL 4600: Standard for Safety – Evaluation of Autonomous Products.

Ward, F. R. and I. Habli (2020). An Assurance Case Pattern for the Interpretability of Machine Learning in Safety-Critical Systems. In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, Cham, pp. 395–407. Springer International Publishing.

Werling, M., R. Faller, W. Betz, and D. Straub (2025). Safety integrity framework for automated driving.