The Missing Link to Safe AI Homologation

Mark Locherer
DHBW Ravensburg
Friedrichshafen, Germany
University of Siegen
Siegen, Germany
locherer@dhbw-ravensburg.de

Michael Kordovan
ifm electronic gmbh
Tettnang, Germany
michael.kordovan@ifm.com

Bernd Buxbaum
ifm electronic gmbh
Essen, Germany
University of Siegen
Siegen, Germany
bernd.buxbaum@ifm.com

Thorsten Kever
DHBW Ravensburg
Friedrichshafen, Germany
kever@dhbw-ravensburg.de

Abstract

The release of the EU AI Act and Machinery Regulation provides the necessary foundation for the use of artificial intelligence in safety components. In this context, both regulations define specific requirements for quality management, risk management and technical documentation, mandatory for conformity assessment by a notified body. In this manuscript, we detail the challenges of AI-based safety-related systems during development and certification, as traditional safety assurance approaches fall short. Further, we illustrate the possible homologation procedures and the interplay of both legal instruments. Additionally, we outline the notion of the efficient assurance argument, including its purpose in certification, to provide a means for quick system assessment and its characteristics during development, in particularly, in designing and maintaining the system in meeting the required tolerable risk level during its lifetime. Based on this, we detail the specific regulatory requirements from both regulations and analyze the characteristics of the efficient assurance argument in meeting these qualities. We conclude, that this type of assurance argument can be seen as the missing link to safe AI homologation, since it serves the engineering team during risk management, system design, verification, validation and testing and operations, as well as the notified body for type approval by demonstrating regulatory compliance.

1 Introduction

So far safety-critical systems (SCSs) have been based on hardware (HW) or software (SW). With the introduction of the Machinery Regulation (MR) and the AI Act (AIA) in the European Union (EU), the legislative foundation for artificial intelligence (AI) in general, but more specifically in safety related contexts, has been build. This offers new perspectives and possibilities for the usage of AI in safety-critical applications. Meanwhile, examples can be observed in our everyday lives, such as autonomous driving [1, 2], or the analysis of medical images in the healthcare domain [3], among others [4].

Article presented at the Second IASEAI Conference (IASEAI'26).

Nevertheless, the utilization of AI-based SW comes with several implications. Most importantly, the "semantic gap," which is the lack of a specification detailing a concrete input-output relationship. Moreover, in safety critical contexts, such as autonomous decision making, its application is challenged by the "responsibility gap," i.e., the inability to hold someone responsible for dangerous outcomes, and the "liability gap," which stands for the difficulty of assigning legal consequences to responsible entities [5]. In this regard, regulation and standardization can form an important step towards shedding light on these concerns by providing well-structured rules and principles. The semantic gap emphasizes the importance of new directions in safe system development, i.e., assurance driven development (ADD) of AI-based SCSs, since traditional HW and SW-based standardization are only partially applicable.

In this manuscript, we address the following aspects: (1) We depict the challenges in the assessment of AI-based safety components. (2) Subsequently, we outline the general homologation procedure for AI-based safety components used in the industrial machinery sector according to MR and AIA in the EU. (3) In this context, we illustrate the efficient assurance argument (AA), its characteristics, and how it integrates within an ADD lifecycle process. (4) Additionally, we analyze the relationship between regulatory compliance and the efficient AA and how this serves in reaching compliance and type approval. (5) Lastly, we provide some ideas on how type-approval by notified bodies can be conducted and related challenges in implementing imposed regulatory requirements. In this context, we claim that—due to the challenges related to AI systems—efficient AAs can be seen as the "missing link" to conformity assessment through a notified body.

This paper is organized as follows: In Sec. 2 we list important terms, provide an introduction to AAs and the challenges related to AI system verification and validation (V & V). In Sec. 3 we detail conformity assessment according to EU regulation. The efficient AA is illustrated in Sec. 4. Subsequently the discussion follows in Sec. 5. Finally, we conclude in Sec. 6 by providing a brief summary of our analysis.

2 Background and related works

A "safety component" is part of a system ensuring environment, health and safety (EHS) [cf. 6, Art. 3(14)]. In this paper, we refer to systems containing safety components as SCSs. As emphasized in [7], a safety component, in the context of the MR, only relates to machinery [cf. 8, Art. 3(3)]. The terms risk and safety are closely related. "[R]isk" is defined as the likelihood of the combination of harm occurring and its severity [9] and "safety" as "freedom from risk which is not tolerable" [cf. 10, Cl. 3.14], which can be achieved with a "safety function," aiming at risk removal or mitigation [cf. 8, Art. 3(4)]. Functional safety is defined as the fraction of safety in a given system that depends on the proper functionality of the safety component and additional risk mitigation approaches [cf. 11, Cl. 3].

Classical safety standardization defines the safety-integrity level as the central element, which determines, the necessary amount of risk reduction of a safety function based on certain criteria and measures and rigor to implement it, capable of reducing this amount of risk.

For "AI system[s]," generating outputs from inputs for a specific task, [cf. 6, Art. 3(1)], the mere ticking-off requirements in wide lists is insufficient, hence, AAs are discussed in [12], to demonstrate that selected AI-based technology, given a specific use case, is sufficient, instead of following prescribed rules, as it is with classical SW in current standardization, [cf. 13–16].

Prior, to "the first making available of a product" on the EU market [6, Art. 3(12)], SCSs require "conformity assessment," which is the procedure for high-risk AI systems demonstrating compliance to regulation [6, Ch. III, Sec. 2] [cf. 6, Art. 3(20)].

In the context of regulation, an analysis of the AIA and trustworthiness is given in [17]. A policy research on the interplay between MR and AIA for AI-based SCSs is provided in [7].

2.1 Challenges of safe AI systems

Classical safety development is based on the V-model paradigm, which makes use of a clear input-output relationship in the system specification, i.e., each requirement in the specification can be linked to a particular system part, such as SW or HW [18–20]. For AI-based SCSs the semantic gap [5, 21] challenges this clear traceability [22] and is further aggravated by the black-box nature of

those systems, as the dedicated system behavior is non-separable from the rest [cf. 23, Cl. 9.2]. Thus, V & V of AI-based systems, mandatory for system homologation, faces many challenges.

Furthermore, the correctness of predictive results given by an AI model is largely governed by uncertainty, i.e., the inability to state, whether given predictions are valid. With this regard, uncertainty is associated with complexities related to environment, observation and the system itself. To describe this distinction, an "onion-layer model" has been proposed in [24], and the "manifestations of uncertainty" in [25]. This can be reflected in the following example: The environment is highly diverse and complex, e.g., weather conditions or traffic participants. A limited technical system, e.g., a sensor, is used to perceive it, potentially resulting in partial environment recognition (observation), due to technical limitations, such as noise. The AI system, with its internal deficiencies, uses this information to fulfill its task. This uncertainty leads to "assurance uncertainty," that characterizes reduced AA trust [cf. 25, Sec. 3.2].

2.2 Assurance argument

To cope with these challenges, assurance approaches for AI-based SCSs, as discussed in [12, 25–27] have been proposed. The first standards on the intersection of AI and safety, i.e., ANSI/UL 4600, ISO/PAS 8800 and ISO/IEC TR 5469, mandate the compilation of an AA [28].

A "safety case" is defined, as "[a] structured argument, supported by a body of evidence that provides a compelling, comprehensible and valid case that a [system] is safe for a given application in a given environment [. . .]" [cf. 29, Sec. 3]. We use the term assurance argument (AA) as a synonym for safety case throughout this paper.

An AA consists at least of (a) the claims, which are assertions about safety or system properties, assumed to be true, (b) the argument, which is a supportive element, why a given claim is regarded as true, and (c) the evidence, that substantiates the argument with details. Therefore, an AA can be seen as a decomposition of an initial top-level claim, stating that a system is acceptably safe, into several sub-claims with the intention of being able to provide a conclusive set of evidence to any claim after a sufficient number of divisions, i.e., in which a claim can be instantly confirmed by evidence. Moreover, an explanatory statement has to be provided for the decomposition structure from high-level claims into more specific claims [cf. 30, Sec. 5]. The claims have to cover all requirements with the intention to show that the SW behavior aligns with its objective and ensures system safety [cf. 31, Ch. 1]. An AA skeleton for an AI system can be found in [32, Annex B].

Relevant information to populate the AA must be derived from the AI component's task and environment. This is given as a mapping of requirements from the broader system level context to the AI component level. While task and environment play a crucial role for assumptions, justification and context in the AA, the task, i.e., its requirements inform the selection of goals, strategies and solutions. Requirement assurance, i.e., safety assurance, can be provided via evidence such as quantitative safety performance indicators (SPIs).

3 Conformity assessment

Conformity assessment for safety components according to MR or AIA, as detailed below, requires a notified body, i.e., an external third-party authority that conducts conformity assessment according to a particular regulation, such as the AIA or other applicable legislation [cf. 6, Art. 3(22)]. The authors of [7] mention a potential shortage in notified bodies for AI-based SCSs. The notifying authority, which is a national entity, is firstly required to assess, designate and to notify "conformity assessment bodies" [6, Art. 3(19)], i.e., before a notified body can operate in the field of AI and functional safety (FuSa), it has to be authorized by such a notifying authority.

Machinery Regulation The Regulation (EU) 2023/1230 on machinery safety requirements (MR) [8] as the successor of Directive 2006/42/EC [33] specifies the requirements on machinery in the context of product safety and design starting from January 20th, 2027. In Art. 20 it is pointed out, that a "product [. . .] which is in conformity with harmonised [sic] standards or parts thereof the references of which have been published in the Official Journal of the European Union shall be presumed to be in conformity with the essential health and safety requirements [. . .]." Therefore, when harmonized standards are utilized in product development, the product is assumed to comply with the regulation.

However, specific products, such as SCSs, have unique requirements. According to [8, Sent. 54]: "[T]he conformity assessment of a safety component or a system with self-evolving behaviour [sic] ensuring safety functions should be carried out by a third party, [. . .]" This means that a notified body as a third party is required for conformity assessment, i.e., the homologation, for "safety components" or systems demonstrating self-evolving behavior, even though the term "self-evolving behavior" is neither clarified in the MR nor in the AIA [cf. 7]. Nevertheless, Art. 25 specifies different "conformity assessment procedures for machinery and related products," which require a notified body. These are illustrated in Figure 1 and characterized as follows:

a) "EU type-examination" (Module B): Conformity assessment procedure for continuous production: A notified body assesses the technical implementation of the product by investigating a sample "type" that resembles the SCS intended for market placement and its documentation [8, Annex VII]. This is followed by "internal production control" (Module C), in which the manufacturer testifies that the "manufacturing process and its monitoring" are suitable for ensuring ongoing production consistency and quality [8, Annex VIII].
b) "[C]onformity based on full quality assurance" (Module H): The notified body assesses the quality management system (QMS) used by the manufacturer to assure that the production meets regulatory compliance [8, Annex IX].
c) "[C]onformity based on unit verification" (Module G): The individual product, intended for market placement, i.e., the unit, is examined by a notified body. The manufacturer produces a technical documentation depicting conformity to EHS requirements, which is used for assessment [8, Annex X].

This is followed by the relevant steps in the AIA, as noted under Para. 3. Finally, the EU declaration of conformity stating that the requirements on EHS are fulfilled is created and the CE marking is affixed to the equipment.

Noteworthy, as pointed out in Annex I(A)(5, 6), the individual assessment through a notified body is applicable to safety components that make use of machine learning (ML)-methods. After positive assessment through a notified body the machinery or equipment can be placed on the EU market by the manufacturer as this implies that EHS requirements according to Art. 8 are fulfilled [8].

AI Act This concern is reflected in Regulation (EU) 2024/1689 (AIA) as well, where AI systems are categorized based on their associated levels of risk. SCSs are classified as high-risk (cf. Art. 6), for which, as specified in Art. 43, conformity assessment conducted by a notified body is mandatory. An audit includes the investigation of the technical documentation, demonstrating compliance to the specific requirements outlined in Arts. 9 – 15, such as a risk management system (RMS) and data governance, and the implementation of a QMS in Art. 17 [6].

The AIA mandates the application of the conformity assessment of the relevant act for the given high-risk AI system [cf. 6, Art. 43(3)], i.e., in the case of a safety component developed for machinery equipment, the MR is applicable. Additionally, the AIA requires the conformity assessment by a notified body according to points 4.3 – 4.5 and 4.6(5) in [6, Annex VII], which considers the technical documentation including dataset access, additional evidence generation, model access and potential retraining in case of insufficiencies, respectively.

Independent, of the conformity assessment procedure selection in the MR (cf. Fig. 1), the implementation of a QMS is mandatory, according to the AIA [cf. 6, Art. 17]. In the case of the technical documentation and the QMS, the existing technical documentation must be extended by the specifics from the AIA according to Art. 11 [cf. 6, Art. 11(2)] and, if given, the existing QMS must be extended by the requirements under Art. 17 of the AIA [cf. 6, Art. 17(3)].

The complete homologation procedure for a safety component according to the MR and AIA is depicted in Fig. 1.

4 Efficient assurance arguments

For the purpose of homologation, we see great potential for the efficient AA, whose key characteristics are based on the "efficient argument," detailed in [31] for classical SW, and discussed in [34], in the context of automotive AI-based safety. From a notified body perspective an AA has to demonstrate the following properties: 1. The assessor can ascertain with ease that system risk is tolerable, and

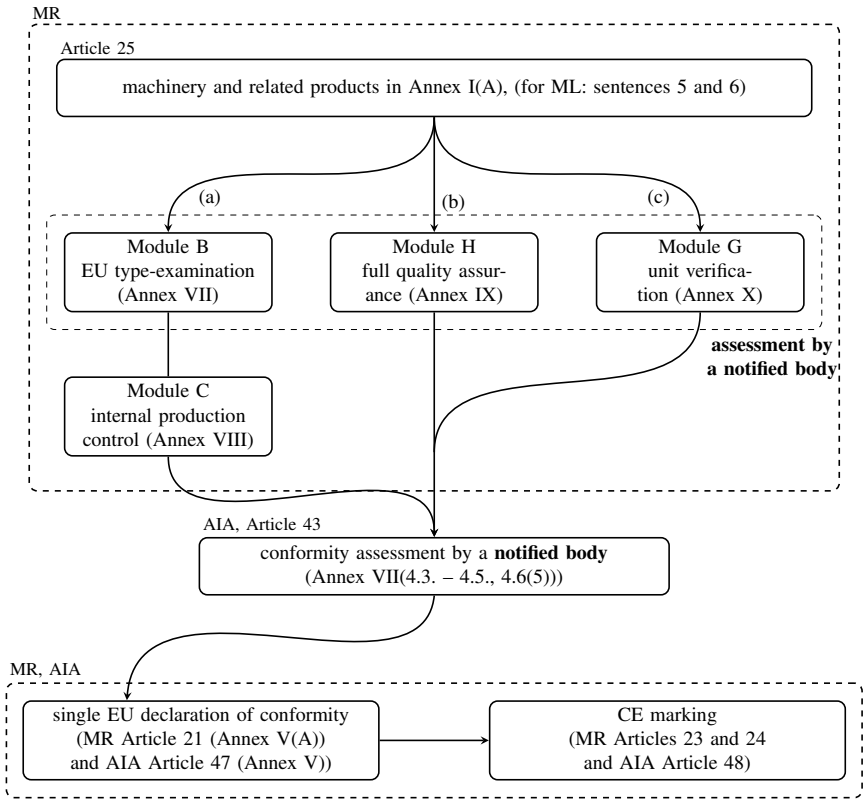

Figure 1: Conformity assessment routes for machinery and related products acc. to Article 25(2) of the MR and the AIA.

hence, fit for purpose and achieving FuSa, without cumbersome complete system testing. 2. The determination of whether the AA suffices for a proper pre-deployment safety case is achieved by investigating, testing, and scrutinizing claims, arguments, and evidence in isolation. 3. Easily detect limitations in argumentative structures and evidence depiction. 4. To assure with confidence, based on a flawless AA, that the system is certifiable [31, 35]. Even in the case of traditional safe SW development this is challenging [cf. 31, Ch. 1].

This leads to the characteristics of an efficient AA: (a) "Assurance driven development and reusable artifacts," which comprises an implemented ADD approach and the application of other tooling and methodologies, which are well-known and qualified. (b) "Precise claims," which help to transform a specific safety requirement into quantifiable AI model property. (c) "Validatable models and assumptions," that concern the understandability of AI model, environment and problem goal. (d) "Rigorous chain of reasoning and evidence," which constitutes of the potential, to assess provided evidence, and the decomposition structure of the arguments.

In this context, Item (a) helps a notified body to determine, whether the assessment properties of an AA can be achieved in first place. An ADD methodology, as shown in Fig. 2, is paramount in serving as the foundation for the compilation of a holistic AA, as well as, the usage of qualified tooling and methods, such as specific SW libraries and HW platforms used to develop and operate AI models.

In case of an AI-based safety component, safety requirements are passed down from the overall system level, to the AI technology level (the model). These requirements have to be addressed in the AA, in which the claims, argument and evidence structure has to be defined according to Items (b) – (d). Precise claims (Item (b)) relate to this transition. They help to specify these requirements in a measurable way, meaning that the derived AI safety requirements become a quantifiable predicate of the utilized AI technology. As noted in[36–38], this poses a great challenge.

Therefore, the understanding of task and environment helps to link to the selected AI technology. The existence of specific models, suitable for given tasks and environments give a notified body the chance to understand what was done and how it was done (Item (c)). Using complex and potentially unknown AI technology challenges this requirement. Finally, Item (d) refers to a proper AA decomposition strategy, reflecting the potential of whether the AA is compelling and the provided evidences, deliver appropriate risk estimates for a specific claims.

When transitioning to autonomous operation, SPIs, as outlined in [30, 39, Ch. 16] and [40], can serve during the complete system lifetime as AA validity monitors, i.e., checking the claims for validity (Items (b) and (c)). This is based on the reality, that even after diligent design and testing, some risks will remain. Therefore, there is a chance to track AA violations before critical incidents happen, by using SPIs, whose feedback serves as input for system updates. Instead of measuring a property, as a key performance indicator (KPI)[1] would do, such as the true positive rate, the SPI checks AA validity. In the case of a person detection system a SPI could measure the frequency of false negative person detections in successive frames and indicate AA validity violation if more than a certain threshold, e.g., five subsequent false negatives, have been detected, i.e., it acts as a monitor for a specific claim. In the case of violation, root cause analysis or safety analysis, as outlined in [32, Cl. 13], has to be conducted, to find potential gaps, leading to system updates.

A simplified ADD lifecycle process (Item (a)) and how SPIs serve as measures (cf. Item (b)) to keep the AA valid is shown in Fig. 2, which is based on [32, 40]. In traditional safety engineering, during design, hazard analysis and risk assessment (HARA) is conducted to identify all relevant hazards, derive the risk and apply mitigation measures [e.g., 41, 42]. Nevertheless, some hazards remain undiscovered, or newly emerge due to drift. In assisted systems the operating context changes to an open environment and likewise, additional uncertainties are introduced, potentially leading to new hazards. This requires iterative V & V and testing in the designated environment to uncover hidden hazards, until the necessary safety target level for deployment is reached. This aspect is covered by the notion of "triggering condition[s]" in [43, Cl. 4]. Hence, there is a feedback from the V & V stage to design, in which adequate measures are implemented, considering the triggering conditions to reach a tolerable risk level. Nevertheless, not all triggering conditions will be identified during V & V and testing, although, the objective of engineering is to achieve optimal system safety. Ultimately, in assisted systems, a human operator is assumed to be responsible for ensuring system safety, closing the gaps of responsibility and liability. This is reflected by the monitoring component during operations. Nevertheless, not all of the uncovered aspects can be addressed by a human operator, leading to system recalls and updates due to technical limitations, even though, these were not initially intended. Examples for ADD processes can be found in ISO/PAS 8800 [cf. 32, Fig. 7-1], and [27, 44–46].

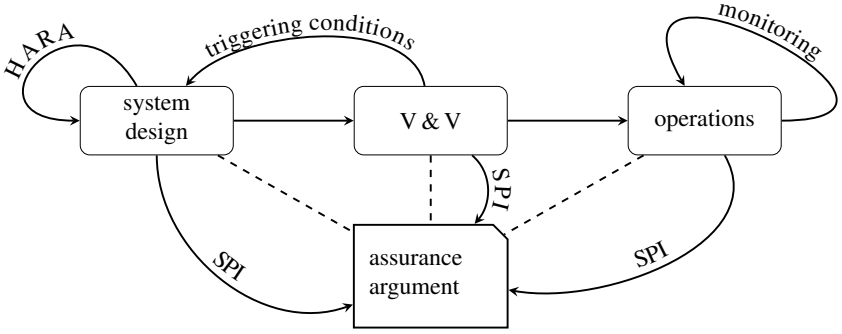

Figure 2: A simplified ADD lifecycle process, incorporating design, V & V and operations. The three stages are connected to the AA with dashed lines to indicate, that prior deployment, the AA is based on these stages. However, to monitor the validity of the AA, SPIs are used and connect the individual stages post deployment.

[1]The term "safety related KPI" is used in ISO/PAS 8800 in the sense of SPI [32, Cl. 3.2.10].

5 Discussion

Despite the shift from the Directive on machinery to the MR and the release of the AIA, highlighting systems with autonomous behavior and ML, there is a current lack of concrete, systematic approaches for conformity assessment accepted by notified bodies, unlike those available for traditional SW development [12]. As highlighted in [17], the absence of harmonized standards complicates the homologation process. Nonetheless, as argued in [17], the AIA lays the groundwork for AI governance, establishing general requirements such as the implementation of a QMS, which must define methods to ensure the system's quality and performance. Additionally, a RMS is necessary, starting with identifying known and foreseeable risks to highlight potential system weaknesses, referred to as "trustworthiness analysis" in [17]. Risk assessment then analyzes these risks under normal and misuse conditions.

Nomenclature Resolving the ambiguity in the notion of "fully or partially self-evolving behaviour [sic]" in the MR [8, Annex I(A)], interpretable as either SW that was obtained via ML (self-evolvement during training), or as SW which continuously adapts even after deployment (ongoing self-evolvement), is mandatory in order to assign the correct category of machinery in the MR. Further, it should be clarified how this definition relates to that of the "AI system" in the AIA [cf. 6, Art. 3(1)]. Nevertheless, the decision that an AI system is considered as "high-risk" must be based on whether the given product is listed under the harmonization legislation under [6, Annex 1] or is a safety-component falling under this list and whether the applicable legislation (in our case the MR) mandates third party conformity assessment [7].

Homologation During homologation, a system might be checked against predefined testing principles. A requirement from such a testing principle could be the detection of a black cylinder to verify the necessary sensor detection capabilities [cf. 47, Sec. 3.3.3]. For AI-based SCSs this has several implications: 1. The challenges outlined in Sec. 2.1 impede the applications of the V-model. 2. Some of the predefined testing principles—according to which an assessment is conducted—might be easily learnable by AI algorithms, hence, a regular testing catalog could be exactly memorized and recalled under testing conditions, thus, rendering this approach void. This potentially results in high certification cost, due to the inability of testing against clear specification with limited boundary conditions. 3. Structural challenges, related to the notification of conformity assessment bodies [7] and their potential to assess AI-based safety components. 4. The current lack of detailed approaches to AI system development in current standardization [28, 48], particularly related to providing sound evidence in the AA [25, 36], exacerbate the situation.

The application of AAs is stated in neither the MR, nor the AIA [6, 8, 12]. We see great potential in using efficient AAs for homologation, as given per Sec. 4. Such an AA protocol, for which an ADD process and reusable artifacts are required, allows for precise claims, validatable models and assumptions and a rigorous chain of evidence, that is required if certification cost reduction with increased rigor lie in the focus. Hence, this approach can be seen as the missing link to AI-based SCS homologation.

To make an efficient AA possible, i.e., one that can be evaluated by an assessor with limited resources (Item (d)), it is important to have a proper ADD lifecycle—as a prerequisite for the fulfillment of the presented characteristics—in place (Item (a)). Moreover, the development either within or among organizations of a sound basis of reusable artifacts is mandatory. This could be achieved by compiling a catalog, in which information on AI technology used for particular task and environment combinations is collected, i.e., which AI model performs well on a particular task and environment. Moreover, a proper understanding of safety-concerns and risks, existing in general and distinct setups has to be collected [cf. 49]. This helps building an understanding of models and assumptions (Items (a) and (c)).

If these reusable artifacts lack for general combinations of model, task and environment, it might be possible to determine assumptions for particular problem subdomains. This seems reasonable considering models constructed by a decision tree algorithm and foundation models. It is important to take AI technology into consideration, that is well-used and understood, which means, that starting with simple models and compact neural networks is encouraged, as these are more easily understandable and techniques for their V & V might be in place [50, 51] (cf. Items (c)). Formal

methods can potentially demonstrate clear system boundaries in neural networks, required for evidence collection in precise claims (Item (b)).

To initiate the process, it is important to experiment with AI technology and feed an artifact database with already gained experience or on the basis of HARA or other alternative techniques. The usage of particular SPIs for specific claims can also be part of such a database (Items (a) and (b)). In this context, standardization can also serve as way to share reusable artifacts and working assumptions (Items (a) and (c)). Nevertheless, this topic is still a great challenge is therefore, an ongoing field of current research and development, as indicted by [12, 34, 37, 38, 52].

5.1 Regulation

In traditional FuSa, the requirements for a technical documentation and a QMS are well established in its management. For instance, the IEC 61508 series requires a QMS (cf. [41, Cl. 6] and [13, Annex F]) and specifies a "hazard and risk analysis," cf. HARA, as outlined in [41, Cl. 7]. Similarly, ISO 26262 mandates a QMS, such as ISO 9001 [53] and IATF 16949 [54], and requires HARA as detailed in ISO 26262-3 [42, Cl. 6].

Nevertheless, the mere presence of these processes does not imply that they cope with the particularities of AI systems. Hence, the specifics outlined in the AIA and the MR have to be incorporated into the existing methodologies. ETSI TR 104 119 provides additional guidelines on the establishment of the AIA regulatory requirements [55].

In the following we outline, how the particular requirements given by the MR and AIA can be addressed with the compilation of an efficient AA, as specified in Sec. 4. However, our focus is on FuSa, and thus, we only consider those requirements in the evaluation, as this is the primary objective of this paper. Besides that, non-safety aspects can be addressed as well in similar arguments, which might require different means for their management, such as fairness [56, 57], as indicated in [6, Art. 9(9)].

Machinery Regulation

Technical documentation The MR mandates the compilation of a technical documentation according to [8, Annex IV]. It has to contain the following information, namely (a) applied methods for risk assessment, EHS requirements and mitigation measures, (b) machinery conformity verification reports, (c) applied methods to meet specification requirements, (d) measures to ensure ongoing regulatory compliance, (e) if applicable, disclosure of safety-relevant SW or logic, (f) in case of data–driven safety-related controls, a description of capabilities, limitations relevant during the lifecycle process, and (g) test results indicating potential safe assembly.

The ADD lifecycle (Item (a)) details the pathway from requirements, over HARA, design, V & V and operations to not only ensure proper engineering, but also demonstrate how requirements and specification are met. In case the safety-relevant logic has to be disclosed, the AA decomposition depicts how requirements are met and the logic behind it (Item (d)). This might also help in protecting intellectual property, as it could be clear that the disclosure of limited parts is sufficient. In case of AI-based safety components, the AA considers all abilities and restrictions (Items (b) and (c)) under which the system functions. Furthermore, the V & V and testing stage ensures that a necessary level of risk mitigation is reached, also given by the monitoring of precise claims and its evidence in Items (b) and (d). The evidences created during the ADD process can assist in compiling the documentation.

AI Act

Technical documentation The sofar MR-based technical documentation has to be extended for SCSs (high-risk AI systems) to detail the realization of Arts. 9 – 15, according to [6, Art. 11]. A key requirement is that it must reflect the current status of the AI system, even after deployment. Further requirements consider descriptions of (a) the AI system itself, comprising purpose, integration within the broader system context, (b) the system elements and applied methods, including algorithmic, data requirements, V & V and testing procedures, etc., (c) monitoring and functioning, detailing abilities and limitations, (d) selection of performance metrics, (e) the RMS, and (f) lifecycle modifications, as stated in [6, Annex IV].

The RMS is considered as a "continuous iterative process planned and run throughout the entire lifecycle [. . .]," that must be able to identify and analyze the "known and the reasonably foreseeable risks," emerging risks, and other potential risks, within "its intended purpose, and under conditions of reasonably foreseeable misuse" utilizing a "post-market monitoring system" [cf. 6, Art. 9(2)]. Moreover, the risk mitigation or control measures shall result in an overall tolerable risk level [cf. 6, Art. 9(5)]. Testing has the goal to identify the "most appropriate and targeted risk management measures," for the AI system to "perform consistently for their intended purpose [. . .]," moreover, "[t]esting shall be carried out against prior defined metrics and probabilistic thresholds that are appropriate to the intended purpose" [cf. 6, Art. 9(6 – 8)].

In this regard, precise claims, as given by Item (b), not only detail the exact purpose of the AI-based safety function, but also, how the requirements from the broader system context relate to the AI system. Further, assumptions and context, shared along the definition of claims, detail the environment, SW and HW, as well as capabilities and limitations, relevant for a given feature. In this context, it is important to consider Items (a) and (c), to ensure, that assumptions hold and are understandable for a notified body. By decomposing top-level claims into sub-claims, immediately supportable by evidence, the applicability of these claims is given if the assigned metrics function, which relates to Items (b) and (d). The evidences and artifacts generated during the ADD process, past deployment, can be used to automatically update the documentation. In summary, the AA reflects the actual state of the system and documents its safety-critical aspects.

Considering the RMS requirement in [6, Art. 9], an ADD maps directly to it (cf. Item (a)), as a necessity for compiling an efficient AA. It must ensure that the set of safety requirements mapped to the AI system are fulfilled during design, V & V and after deployment. Hence, we can state that a AA requires a RMS, i.e., risk management is continuously conducted in ADD, and thus, forms the foundation for regulatory compliance. Moreover, the post-market monitoring system, as outlined in [6, Art. 72], is also mandatory for an efficient AA, and as noted monitors, such as SPIs, can assist the compilation of precise claims and corresponding evidence (cf. Items (b) and (d)), and serve as a feedback mechanism, posterior deployment, for ongoing AA validity.

Quality management system A QMS is mandated in [6, Art. 17], to ensure regulatory compliance. This includes procedures for (a) design and development, (b) quality control, including V & V and testing during development, (c) data management according to Art. 10, ensuring quality, (d) the RMS, particularly for conformity assessment and system management, (e) the post-market monitoring system, for ongoing monitoring and incident reporting, (f) record-keeping according to Art. 12, guaranteeing appropriate operation with respect to the intended function.

In this sense, Item (a) with an ADD process serves as the basis for a QMS. It considers the complete system lifecycle, including the interactions of the individual stages and how these affect the compilation of evidences and artifacts for the AA. Data management, given by the "[d]ataset lifecycle model," shown in [32, Fig. 11-3], is also integrated. The relevant considerations for the RMS and post-market monitoring system are sketched in Para. 5.1. Further, the definition of precise claims and evidence generation (Items (b) and (d)) implement the requirements for record-keeping, as each claim must be measurable, using tools like SPIs. Additionally, operationalization is given by defining SPIs as evidences for claims to ensure system quality by on-going monitoring. Their definition, including thresholds and resolution protocols, is goal dependent (derived from system task and environment). Collectively, SPIs can be seen as part of the post-market monitoring plan mandated in [6, Art. 72].

Summary The regulatory requirements, such as the QMS, and the joint technical documentation, are mandatory for AI-based SCS assessment. This can also be seen in the light of methods under reusable artifacts (Item (a)) that must be in place when compiling an efficient AA. Hence, in conformity assessment the presence of these tools, their capabilities and interaction only allows for the compilation of a structured argument. In other words, a proper ADD lifecycle process that covers all these aspects is required and to argue, whether Items (b) and (d) are possible in first place, for which the methods under Item (a) must be implemented. Although the AIA terms AI-based SCSs as "high risk," they should not be high risk systems anymore, since the compilation of the AA, as detailed in Sec. 4, aims at removing exactly these risks, hence, the term "CE-marked AI system" is preferred by [7].

In the context of homologation, the efficient AA, given the characteristics depicted under Sec. 4, can be seen as the "missing link" providing firstly, a tool to fulfill the requirements under the MR and

AIA for the manufacturer and secondly, an effective and comprehensible means to assess conformity by notified bodies. Obviously, an ADD method has to be in place to be able to create and maintain an AA during the different lifecycle phases of HARA, design, V & V and post deployment.

Besides check-listing that the correct methods (Item (a)) are used, possible approaches to homologation might be the close supervision of projects during development, or the creation of unknown but representative operational datasets according to specification by the certification authority used for type-approval. Further approaches might be the definition of well-understood models, as well as approved and trusted methods in the sense of Items (a) and (c) in testing principles.

The resulting workload for developers and assessors must be manageable in terms of know-how and workforce. An assumption for Item (a) is given by the qualification and later cost amortization of reusable tools and methods in the long run [31]. While this approach is likely suitable for large companies, smaller ones might face serious problems in the establishment of qualified tools and processes, due to cost and labor. Further, the industrial industry sector differs much from the automotive sector, and hence, the implementation of monitoring systems over the complete lifecycle not only affects the manufacturer of AI components but also potential clients. An overarching infrastructure, ensuring AA validity, has to be implemented for system monitoring and potential update mechanisms, which might not always be possible or desirable.

This also leads to the discussion of responsibilities assigned to the ADD process which is also emphasized under "[s]afety culture," in [39, Cl. 5]. A proposal for roles is presented in [58] for SCSs incorporating AI, and specifically for the ML Operations process in [59], being the core in data-based ADD. Different areas of expertise have to collaborate extensively, such as systems engineering, safety engineering, domain experts, requirements engineering and the AI-related responsibilities, e.g., AI experts and data scientists. Further, external assessors share common touch points.

6 Conclusion

In this paper we provide an overview of the homologation process of AI-based safety components according to MR and AIA, for which a notified body is mandatory.

The challenges in this process are specifically related to the nature of AI systems, their V & V and uncertainty handling. For classical SW engineering, the notion of the efficient AA was introduced, to characterize an argument that details safety and makes external assessment possible. The characteristics of such an argument are "assurance driven development and reusable artifacts," "precise claims," "validatable models and assumptions" and a "rigorous chain of reasoning and evidence."

In our study, we characterize the efficient AA as an argument that provides the necessities for third party assessment. Further, we analyze the requirements from the regulations and compare them with the characteristics of the efficient AA and propose how they can meet regulatory compliance criteria. We discover a great overlap between these requirements, such as documentation management and QMS, and the efficient AA. Lastly, we outline possible ideas to certification and challenges related to implementing ADD, particularly for smaller companies. Hence, the efficient AA can be seen as the missing link to homologation.

The definition of monitoring mechanisms for particular claims, such as SPIs, AA sufficiency, the compilation of a database with reusable artifacts detailing use cases in combination with AI technology, risks and safety assurance techniques have great potential for further research. In our point of view, this is a mandatory requirement to make an efficient AA and conformity assessment possible.

Acknowledgments and Disclosure of Funding

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Nick Webb, Dan Smith, Christopher Ludwick, Trent Victor, Qi Hommes, Francesca Favaro, George Ivanov, and Tom Daniel. Waymo's Safety Methodologies and Safety Readiness Determinations, 2020.

[2] Matthew Schwall, Tom Daniel, Trent Victor, Francesca Favaro, and Henning Hohnhold. Waymo Public Road Safety Performance Data, 2020.

[3] Rogier R. Wildeboer, Ruud J. G. van Sloun, Hessel Wijkstra, and Massimo Mischi. Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods. *Computer Methods and Programs in Biomedicine*, 189:105316, 2020.

[4] ISO/IEC. ISO/IEC TR 24030: Artificial Intelligence – Use cases, 2024.

[5] Simon Burton, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan, and Zoe Porter. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279:103201, 2020.

[6] European Parliament, Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024.

[7] Håkan Burden, Susanne Stenberg, and Kristian Flink. When AI meets machinery – the role of the notified body. Technical Report 2024:56, 2024.

[8] European Parliament, Council of the European Union. Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery and repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC (Text with EEA relevance), 2027.

[9] DIN. DIN EN ISO 12100: Safety of machinery – General principles for design – Risk assessment and risk reduction, 2010.

[10] ISO/IEC. ISO/IEC GUIDE 51: Safety aspects – Guidelines for their inclusion in standards, 2014.

[11] IEC. IEC 61508-4: Functional safety of electrical/electronic/programmable electronic safetyrelated systems – Part 4: Definitions and abbreviations, 2010.

[12] Rasmus Adler and Michael Klaes. Assurance Cases as Foundation Stone for Auditing AI-Enabled and Autonomous Systems: Workshop Results and Political Recommendations for Action from the ExamAI Project. In *HCI International 2022 – Late Breaking Papers: HCI for Today's Community and Economy*, pages 283–300, Cham, 2022. Springer Nature Switzerland.

[13] IEC. IEC 61508-2: Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 2: Requirements for electrical/electronic/programmable electronic safety-related systems, 2010.

[14] IEC. IEC 61508-3: Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 3: Software requirements, 2010.

[15] ISO. ISO 26262-5: Road vehicles – Functional safety – Part 5: Product development at the hardware level, 2018.

[16] ISO. ISO 26262-6: Road vehicles – Functional safety – Part 6: Product development at the software level, 2018.

[17] André Steimers. Analyse der Vertrauenswürdigkeit von KI im Kontext der KI-Verordnung. In *KI zwischen Ost und West*, pages 1–11. Springer Fachmedien Wiesbaden, Wiesbaden, 2025.

[18] Eric Jenn, Alexandre Albore, Franck Mamalet, Grégory Flandin, Christophe Gabreau, Hervé Delseny, Adrien Gauffriau, Hugues Bonnin, Lucian Alecu, Jérémy Pirard, et al. Identifying challenges to the certification of machine learning for safety critical systems. In *European Congress on Embedded Real Time Systems (ERTS 2020)*, pages 1–10, 2020.

[19] Bernd Spanfelner, Detlev Richter, Susanne Ebel, Ulf Wilhelm, and Carsten Patz. Challenges in applying the ISO 26262 for driver assistance systems. In *Tagung Fahrerassistenz, München*, volume 15, München, 2012.

[20] Alec Banks and Rob Ashmore. Requirements Assurance in Machine Learning. In *Workshop on Artificial Intelligence Safety 2019 Co-Located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, volume 2301, pages 1–4. CEUR-WS.org, 2019.

[21] Carl Bergenhem, Rolf Johansson, Andreas Söderberg, Jonas Nilsson, Jörgen Tryggvesson, Martin Törngren, and Stig Ursing. How to Reach Complete Safety Requirement Refinement for Autonomous Vehicles. In *CARS 2015 - Critical Automotive Applications: Robustness & Safety*, Paris, France, 2015.

[22] Philip Koopman and Michael Wagner. Challenges in Autonomous Vehicle Testing and Validation. *SAE International Journal of Transportation Safety*, 4(1):15–24, 2016.

[23] ISO/IEC. ISO/IEC TR 5469: Artificial Intelligence – Functional safety and AI systems, 2024.

[24] Michael Kläs and Anna Maria Vollmer. Uncertainty in Machine Learning Applications: A Practice-Driven Classification of Uncertainty. In *Computer Safety, Reliability, and Security*, pages 431–438, Cham, 2018. Springer International Publishing.

[25] Simon Burton and Benjamin Herd. Addressing uncertainty in the safety assurance of machine-learning. *Frontiers in Computer Science*, 5:1132580, 2023.

[26] Simon Burton, Lydia Gauerhof, and Christian Heinzemann. Making the Case for Safety of Machine Learning in Highly Automated Driving. In *Computer Safety, Reliability, and Security*, pages 5–16, Cham, 2017. Springer International Publishing.

[27] Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu, and Ibrahim Habli. Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS), 2021.

[28] Mark Locherer, Michael Kordovan, Bernd Buxbaum, and Thorsten Kever. 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, 2025. IEEE.

[29] UK Defence Standardization. DEF STAN 00-056 Part 01 Issue 8: Defence Standard 00-056 Part 01 (Issue 8) – Safety Management Requirements for Defence Systems Part: 01 : Requirements and Guidance, 2023.

[30] Philip Koopman. *The UL 4600 Guidebook: What to Include in an Autonomous Vehicle Safety Case*. Carnegie Mellon University, 2022.

[31] Natarajan Shankar, Devesh Bhatt, Michael Ernst, Minyoung Kim, Srivatsan Varadarajan, Suzanne Millstein, Jorge Navas, Jason Biatek, Huascar Sanchez, Anitha Murugesan, and Hao Ren. DesCert: Design for Certification, 2022.

[32] ISO. ISO/PAS 8800: Road vehicles – Safety and artificial intelligence, 2024.

[33] European Parliament, Council of the European Union. DIRECTIVE 2006/42/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast) (Text with EEA relevance), 2006.

[34] Simon Burton. Safety under uncertainty: Automotive standards for AI safety and research perspectives, 2024.

[35] ACWG. SCSC-141C: Goal Structuring Notation Community Standard, 2021.

[36] Simon Burton, Benjamin Herd, and João-Vitor Zacchi. Uncertainty-Aware Evaluation of Quantitative ML Safety Requirements. In *Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops*, pages 391–404, Cham, 2024. Springer Nature Switzerland.

[37] Benjamin Herd, João-Vitor Zacchi, and Simon Burton. A Deductive Approach to Safety Assurance: Formalising Safety Contracts with Subjective Logic. In *Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops*, pages 213–226, Cham, 2024. Springer Nature Switzerland.

[38] Benjamin Herd and Simon Burton. Can you trust your ML metrics? Using Subjective Logic to determine the true contribution of ML metrics for safety. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 1579–1586, New York, NY, USA, 2024. Association for Computing Machinery.

[39] Underwriter Laboratories. ANSI/UL 4600: Standard for Safety – Evaluation of Autonomous Products, 2019.

[40] Fraunhofer IESE. External Talk: "A Safety Case plus SPIs Metric Approach for Self-Driving Car Safety" (EDCC 2020), 2020.

[41] IEC. IEC 61508-1: Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 1: General requirements, 2010.

[42] ISO. ISO 26262-3: Road vehicles – Functional safety – Part 3: Concept phase, 2018.

[43] ISO. ISO 21448: Road vehicles – Safety of the intended functionality, 2022.

[44] Markus Borg, Jens Henriksson, Kasper Socha, Olof Lennartsson, Elias Sonnsjö Lönegren, Thanh Bui, Piotr Tomaszewski, Sankar Raman Sathyamoorthy, Sebastian Brink, and Mahshid 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.

[45] Marc Zeller, Thomas Waschulzik, Carmen Carlan, Marat Serahlazau, Claus Bahlmann, Zhiliang Wu, Sigurd Spieckermann, Denis Krompass, Simon Geerkens, Christian Siebericchs, Konstantin Kirchheim, Batu Kaan Özen, and Lucia Diez Robles. Continuous Development and Safety Assurance Pipeline for ML-Based Systems in the Railway Domain. In *Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops*, pages 446–459, Cham, 2024. Springer Nature Switzerland.

[46] Marc Zeller, Thomas Waschulzik, Reiner Schmid, and Claus Bahlmann. Toward a safe MLOps process for the continuous development and safety assurance of ML-based systems in the railway domain. *AI and Ethics*, 4(1):123–130, 2024.

[47] DGUV Test, Fachbereich Verkehr und Landschaft and Institut für Arbeitsschutz der Deutschen Gesetzlichen Unfallversicherung (IFA). GS VL 40 – Grundsätze für die Prüfung und Zertifizierung von Rückfahrassistenzsystemen für Nutzfahrzeuge. Technical report, Hamburg and Sankt Augustin, Germany, 2019.

[48] Jon Perez-Cerrolaza, Jaume Abella, Markus Borg, Carlo Donzella, Jesús Cerquides, Francisco J. Cazorla, Cristofer Englund, Markus Tauber, George Nikolakopoulos, and Jose Luis Flores. Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. *ACM Comput. Surv.*, 56(7):176:1–176:40, 2024.

[49] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence, 2024.

[50] Saddek Bensalem, Xiaowei Huang, Wenjie Ruan, Qiyi Tang, Changshun Wu, and Xingyu Zhao. Bridging formal methods and machine learning with model checking and global optimisation. *Journal of Logical and Algebraic Methods in Programming*, 137:100941, 2024.

[51] Saddek Bensalem, Chih-Hong Cheng, Wei Huang, Xiaowei Huang, Changshun Wu, and Xingyu Zhao. What, Indeed, is an Achievable Provable Guarantee for Learning-Enabled Safety-Critical Systems. In *Bridging the Gap Between AI and Reality*, pages 55–76, Cham, 2024. Springer Nature Switzerland.

[52] Christina Klüver, Anneliesa Greisbach, Michael Kindermann, and Bernd Püttmann. A requirements model for AI algorithms in functional safety-critical systems with an explainable self-enforcing network from a developer perspective. *Security and Safety*, 3:2024020, 2024.

[53] DIN. DIN EN ISO 9001: Quality management systems – Requirements (ISO 9001:2015); German and English version EN ISO 9001:2015, 2015.

[54] VDA QMC Qualitäts Management Center im Verband der Automobilindustrie e. V. IATF 16949: Quality management system requirements for automotive production and relevant service parts organisations, 2016.

[55] ETSI. ETSI TR 104 119: Methods for Testing & Specification (MTS); AI Testing; Guidelines for Documentation of AI-enabled Systems, September 2025.

[56] André Steimers and Moritz Schneider. Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health*, 19(6):3641, 2022.

[57] ISO/IEC. ISO/IEC TR 24028: Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence, 2020.

[58] Sangeeta Dey and Seok-Won Lee. Multilayered review of safety approaches for machine learning-based systems in the days of AI. *Journal of Systems and Software*, 176:110941, 2021.

[59] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE access : practical innovations, open solutions*, 11:31866–31879, 2023.