Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components

Michael Kläs, Rasmus Adler, Lisa Jöckel, Janek Groß, Jan Reich
Fraunhofer IESE, Kaiserslautern, Germany
{michael.klaes, rasmus.adler, lisa.joeckel, janek.gross, jan.reich}@iese.fraunhofer.de

Abstract

Artificial Intelligence (AI), particularly current Machine Learning approaches, promises new and innovative solutions also for realizing safety-critical functions. Assurance cases can support the potential certification of such AI applications by providing an assessable, structured argument explaining why safety is achieved. Existing proposals and patterns for structuring the safety argument help to structure safety measures, but guidance for explaining in a concrete use case why the safety measures are actually sufficient is limited. In this paper, we investigate this and other challenges and propose solutions. In particular, we propose considering two complementary types of risk acceptance criteria as assurance objectives and provide, for each objective, a structure for the supporting argument. We illustrate our proposal on an excerpt of an automated guided vehicle use case and close with questions triggering further discussions on how to best use assurance cases in the context of AI certification.

1 Introduction

AI, which in this paper we understand as complex data-driven models provided by Machine Learning (ML), promises improved or additional functionalities that are essential for autonomous systems, e.g., perception for self-driving vehicles. In many cases, such functionalities are safety-critical, so it is highly likely that AI becomes safety-critical as well, meaning that its failure can contribute to accidents. There are already various reports on fatal accidents due to AI-related failures in autonomous vehicles [Pietsch, 2021; Wakabayashi, 2018].

In consequence, regulation [European Commission, 2021] and certification for AI in safety-critical components is being proposed. Regulation and certification are powerful means to prevent the market introduction of unsafe products. This contributes not only to safety but also to the economy as a few unsafe products could affect user acceptance of all similar products. The predictability of legal decisions can thus contribute to economic success as long as liability risk and costs for complying with regulations and standards are not unreasonably high and hinder meaningful innovations.

Unfortunately, existing safety standards are difficult to apply in the context of AI [Salay and Czarnecki, 2018] and revisions are still ongoing [ISO/IEC, 2021]. Therefore, we currently do not have any standards that we can easily apply for certifying AI.

Argument safety claims with assurance cases (ACs) as an established approach in safety engineering may provide an alternative basis for audits and certification in the context of AI [BSI, 2021]. They could structure the arguments for those parts of a solution that are individual and highly innovative. Moreover, they could establish the basis for upcoming evidence-based standards for AI certification.

Initial proposals on how to apply the concept of ACs to AI can be found in the literature. A prominent strategy is to argue the safety objectives and safety requirements [Gauerhof et al., 2020]. As the proposed strategy and patterns abstract from specific safety objectives and derived safety requirements, such approaches also largely abstract from AI-specific safety concerns and required safety measures. Guidance for achieving and arguing safety is thus inherently limited.

One approach for overcoming this limitation is to argue using known AI-related safety concerns and how they are addressed by AI-specific safety measures [Schwalbe et al., 2020]. A disadvantage is that it is hard to argue completeness for the identified and addressed safety concerns. Furthermore, such approaches can not explain yet what safety measures and metrics with the respective thresholds need to be applied to achieve a defined level of safety. To give just one example, neither practical experience nor empirical evidence exists on defining a specific neuron coverage level that would be considered as sufficient when testing a deep neural network for a concrete application.

We think that the concepts and ideas introduced in existing AC proposals can be aligned in a more comprehensible and convincing argumentation if the risk acceptance criteria on which the question of ‘How safe is safe enough?’ is founded, is made explicit in the AC structure itself. We will show that this allows, on the one hand, becoming explicit with respect to AI-specific safety measures and, on the other hand, soundly arguing higher-level safety-objectives.

Contribution. Specifically, we propose using an AC structure that splits ***at an early stage*** into two main claims and related arguments. The first claim refers to the achievement of a probabilistic target value with a certain level of ***confidence*** derived from applying a quantitative risk acceptance criteria. The second claim is that the risk due to “failures” caused by the AI is as low as reasonably practicable due to safety measures applied during the ***AI lifecycle***. In the absence of evidence-based target values for specific safety measures, we propose to ***monitor quality assurance activities*** on a cost-benefit base and define respective stop criteria.

This ensures, on the one hand, that quantitative objectives are explicitly argued and underpinned with evidences. On the other hand, the argumentation over the proposed lifecycle stages contributes to a more comprehensive and justifiable derivation of reasonable safety measures but without the need for predefine targets for specific safety measures. The aim of this paper is to stimulate the discussion about how to argue safety for AI-based functions by rethinking traditional AC patterns and strategies.

Structure. The remainder of this paper is structured as follows: First, we give some background on quality assurance in the context of AI and introduce the concept of ACs as applied in safety engineering (Sec. 2). Next, we discuss existing proposals on how ACs could be used in the context of AI (Sec. 3). Then we introduce an example use case and illustrate our proposal for structuring ACs (Sec. 4). Finally, we discuss a selection of open question (Sec. 5) and conclude the paper with an outlook on possible implications (Sec. 6).

2 Background

2.1 Quality Assurance for AI

AI-based software components raise new challenges for quality assurance due to their functionality being derived from data. Commonly, challenges and safety concerns like lack of specification or interpretability are described [Adler et al., 2019; Ashmore et al., 2019; Felderer and Ramler, 2021; Sämann et al., 2020; Willers et al., 2020]. Several papers collect existing methods and map them to mentioned challenges [Adler et al., 2019; Sämann et al., 2020; Schwalbe and Schels 2019; Willers et al., 2020]. This raises two questions: whether the list of safety concerns is complete, and to which extent the available methods sufficiently address the safety concerns [Adler et al., 2019]. We are currently not aware of any work that could provide a sufficient answer on these questions.

Another approach is to structure possible quality assurance activities and measures according to the phases of the AI lifecycle in which they are applied. Studer et al. [2021] propose, for example, a process model based on CRISP-DM, which is often used in data analysis projects, introducing a quality assurance methodology for each project phase. Ashmore et al. [2019] provide a survey of quality assurance methods generating evidences for key assurance requirements being met in each phase of the AI lifecycle. Here, there is a need to show that the quality assurance methods applied during a phase address all assurance requirements related to this phase, and that the list of assurance requirements is complete.

However, it is difficult to obtain a complete list of quantitative quality assurance requirements. These strongly depend on the task of the AI-based component and its application context. Quality modeling approaches can contribute to a more comprehensive list of quality requirements [Mayr et al., 2012]. Siebert et al. [2021] propose a systematic approach for building such a quality model for a concrete AI-based system that defines the required aspects for each entity of the AI-based system and how they can be measured. Still, further research is needed to better understand (1) to which extent an evidence generated by a certain method contributes to arguing safety, (2) what suitable performance indicators for the evidences are, and (3) when a certain method should be preferred over another for a given context.

2.2 Assurance Cases

ACs are heavily used in practice to assure safety. In particular, if it is very challenging to argue safety, as in the case of autonomous systems. In recent years, standards like UL 4600 [UL, 2021] or reports [Zenzic, 2020] have addressed the development of such AC. The application rule VDE-AR-E 2842-61 [VDE, 2020] already proposes using ACs also for other critical aspects of trustworthiness, such as fairness, as illustrated by Hauer et al. [2021].

An AC is defined as a reasoned, auditable created artifact that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and the underlying evidence and explicit assumptions that support the claim(s) [ISO/IEC/IEEE, 2019].

The left part of Fig. 1 illustrates the three main building blocks of an AC: (1) its top-level claims typically referring to achieved objectives or fulfilled constraints, (2) an argumentation supporting the top-level claims, and (3) evidences on which the argument is based. The right part illustrates the argumentation in a tree structure and its assumptions. The tree is built from reasoning steps that connect lower-level claims with a higher claim that can be concluded from these lower-level claims. If the conclusion is only valid under some assumptions, these assumptions shall be made explicit.

There are different languages for modeling ACs, like the Goals Structuring Notation (GSN) [SCSC, 2018] or Claim Argument Evidence Notation [Adelard LLP, 2021]. The common meta-model of these languages is defined in the Structured Assurance Case Metamodel (SACM) [OMG, 2020]. This paper do not refer to a specific language but focus on the fundamental idea of structuring the argument.

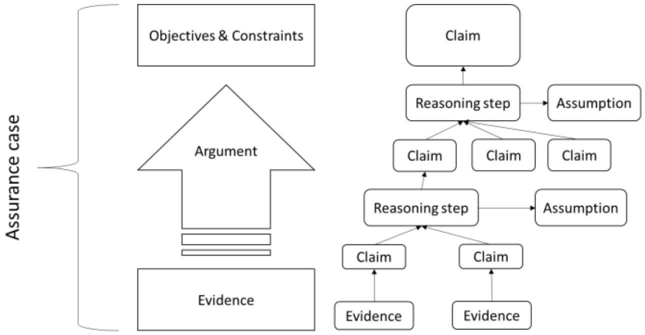


Figure 1: Building blocks and general structure of an AC

3 Related Work

From a safety perspective, ACs are considered a promising approach for arguing safety for AI-based systems, and various authors have already proposed strategies and patterns.

Picardi et al. [2019] presented an AC pattern for ML models in clinical diagnosis systems, which they later refined and supplemented by a process for generating evidences during the ML lifecycle [Picardi et al., 2020]. The activities and desiderata during the ML lifecycle are referred from Ashmore et al. [2019]. The ML assurance claim is argued based on ML safety requirements, operating environment, ML model, development, and test data. In this context, the link between system safety requirements and ML safety requirements is addressed [Gauerhof et al., 2020]. In the recently published AMLAS report, Hawkins et al. [2021] also provide generic argument patterns and a process for ML safety assurance scoping, ML safety requirements, ML data, model learning, model verification, and model deployment.

Wozniak et al. [2020] propose an argument pattern for safety assurance that is aligned with the reasoning for software and hardware in ISO 26262. They argue satisfaction of an ML safety requirement over correctly decomposing the safety requirements into sub-requirements and their satisfaction, appropriate data acquisition, model design, as well as implementation and training of the ML model.

A strategy that does not argue the fulfillment of ML safety requirements is provided by Gauerhof et al. [2018]. They argue that the intended functionality is met by a sufficient reduction of the root causes of functional insufficiencies, which encompass underspecification, semantic and deductive gap.

Based on previous works [Schwalbe and Schels, 2019; 2020], Schwalbe et al. [2020] propose arguing the sufficient absence of risk for deep neural networks (DNN) arising from the insufficiencies they see in their black-box nature, simple performance issues, incorrect internal logic, and instability. They propose a collection of measures to address these insufficiencies, which include V&V as well as best practices during the creation of DNNs and on the system level.

In summary, our review indicates that existing work is driven by the safety community, which adapts established safety patterns and concepts to AI. However, the presented patterns are still on a rather abstract level, and the applicability on a concrete use case comprehensively illustrated from the top-level claim down toward the evidences has not been described yet so far. This might indicate that transferring traditional patterns to AI-based systems proves to be difficult.

We observed two major challenges in argumentation for which existing strategies and patterns still provide insufficient support. (1) Completeness in the refinement of claims in sub-claims appears difficult to show, especially, when approaches argue over the refinement of safety requirements to AI/ML requirements or about addressing ML insufficiencies. For example, if we have a (most likely) incomplete list of insufficiencies, we cannot argue about addressing each insufficiency. (2) Considering the current state of AI quality assurance, the proposed patterns commonly struggle with bridging the gap between a low-level quantitative evidence, e.g., achieving a specific neuron coverage during AI testing, and the claim of sufficient safety for the given application in a convincing manner.

We pinpointed as potential cause of these problems the fact that the risk acceptance criterion underlying the top-level claim on which the argumentation is based is either implicit or different criteria are mixed and are thus not easy to distinguish during refinement. We therefore claim that a clear differentiation will allow more specific argumentation patterns and better attribution of evidences to sub-claims.

4 Building Safety Assurance Cases for AI

In this section, we will first introduce the example we will use to illustrate our concepts. Then we will motivate the consideration of a combination of two risk acceptance criteria to structure ACs for AI. Finally, we will introduce a lifecycle model and use it to argue completeness of the provided refinement.

4.1 Background of the Selected Example

Automated guided vehicles (AGV) are driverless vehicles that transport material. They are used in industrial applications for realizing the flow of material and their safety concepts do not rely on AI [DIN, 2017]. However, their application is limited due to limited understanding of the environment and their safety concept. Autonomous mobile robots (AMR) overcome these disadvantages compared to operator-controlled vehicle by using more sensors and AI. However, the goal of achieving similar performance and flexibility as an operator-controlled vehicle is hard to realize without using AI in safety-critical functions like collision avoidance. Operators of forklifts adapt their speed and safety distance according to various aspects of the persons at risk, including speed, motion path, eye contact, hand gestures signaling right of way, etc. To implement a conservative version of such a human-like collision avoidance system, the AMR needs an AI-based component that understands whether a person at risk has recognized the AMR and gives way to it. A critical failure in this context is that the AMR falsely detects the signaling of right of way. Such safety-critical false detections have to be avoided sufficiently to assure that the AMR drives as least as safe as an operator.

4.2 What does sufficient mean?

The answer to the question of what sufficient means to prevent a safety-critical failure like '*false detections of a human gesture*' depends on the related risks and the risk acceptance criteria, as safety is defined as acceptable risk [IEC, 2010].

We should keep two important aspects in mind when discussing criteria for risk acceptance in settings where AI is part of a safety-critical function: (a) AI is an emerging technology that is still heavily in flux, with unforeseeable developments and improvements in the upcoming years. Thus, coming up with a fixed set of safety measures does not appear to be reasonable. The argument that these safety measures minimize risks as far as reasonably practicable easily becomes invalid. Besides, it would be hard to argue that these measures are as effective as existing ones in safety standards for traditional software. (b) AI is also mainly applied to realize functions that cannot be provided yet by traditional technological solutions.

A risk acceptance criterion that seems reasonable to apply in the context of AI – considering (a) – states that the residual risk after the application of safety measures should be *As Low As Reasonably Practicable* (ALARP). The meaning of 'reasonably practicable' is not static but depends on the state of the technology and the intended application, including the underlying business case and related practical restrictions. Considering ALARP as part of the argumentation assures that when progress in technology allows for safer solutions, we will see progress in safety.

However, doing one's best to avoid and mitigate risks is obviously not enough to argue that the best was sufficient. Accordingly, ALARP is only used in an ALARP region, which is the region between an upper tolerance limit marking unacceptable risk and a lower tolerability limit. Having this in mind is of crucial importance when applying ALARP to AI since the current state of AI technology might not be advanced enough to realize a given application in a sufficiently safe manner. For example, a state-of-the-art traffic sign recognition algorithm might get one of 200 stop signs wrong [INI, 2019]. If used as part of an autonomous vehicle, it may, as a result, regularly ignore someone's right of way. The algorithm might be as good as reasonably practicable but is still not sufficiently safe to be applied in this specific application.

Thus, we need at least a second risk acceptance criterion that gives us a fixed limit.

Most existing products have been developed according to functional safety standards that follow the risk acceptance criterion *Minimum Endogenous Mortality* (MEM). The idea of MEM is that a technical system must not create a significant risk compared to globally existing risks. For example, a product should cause a minimal increase in overall death rates compared to the existing population death rates. This idea leads to very challenging safety requirements and low target failure rates. Depending on the specific task, such low failure rates might be hard to achieve in practice if AI is involved.

An alternative criterion given a fixed target is *Globalement au moins aussi bon* (GAMAB), which says that new technical systems shall be at least as safe as comparable existing ones. However, due to (b) it is hardly applicable in case of many AI-based functions because no technical systems exist yet that provide similar functions.

An approach related to GAMAB is the idea of having a '*positive risk balance*' (PRB). PRB is defined in ISO/TR 4804 as the 'benefit of sufficiently mitigating residual risk of traffic participation due to automated vehicles' together with Note 1 'This includes the expectation that automated vehicles cause less crashes (3.7) on average compared to those made by drivers' [ISO/TR, 2020]. The idea of comparing the new AI-based solution with the existing sociotechnical system can lead to less challenging target failure rates compared to MEM. This opens up new opportunities for arguing safety.

In this paper, we do not discuss how to use this opportunity to derive a target failure rate for an AI-based safety-critical function, as this is very specific for the function and its usage context, but not specific for AI. We do also not discuss how to get from the target failure rate to a target upper boundary on the uncertainty for AI outcomes. Instead, we assume in our example that we would end up with a PRB-derived upper boundary on safety-related uncertainty (u) that we could accept for the AI outcomes: '*The AI must not falsely detect a signal for the right of way that was not actually given in more than one of N cases*'.

Fig. 2 illustrates the relationship between ALARP and a target-based criterion such as MEM, GAMAB, or PRB when providing arguments that an AI-based solution is safe.

ALARP can be considered as requesting a certain *alpha* given by the ratio between the reduction of safety-related uncertainty in the AI outcomes and the required effort/cost. Given the business case for the planned solution and the state of technology, this alpha might vary and is achieved in Fig. 2 at point B. Simply speaking, we request that as long as safety measures exist that would increase safety with reasonable investment, they are carried out. How this rather abstract constraint can be further refined will be discussed in the context of the AI lifecycle presented in Sec. 4.3.

The upper boundary on acceptable safety-related uncertainty u that is derived from the target-based criterion is illustrated in Fig. 2 as a horizontal line. We consequently need to argue that we are confident that the actual safety-related uncertainty is below u. Please note that this is not achieved at point A, but first at C, which we will discuss further, including its implications when talking about testing in Sec. 4.3.

Finally, we will always end up in one of two kinds of situations: a situation where the target-based criterion dominates, i.e., it defines the required investment (cf. Fig. 2), or a situation where ALARP dominates the required investment. An interesting question, which is, however, not directly related to safety, is whether a solution requiring more investment than reasonably practicable should actually be targeted.

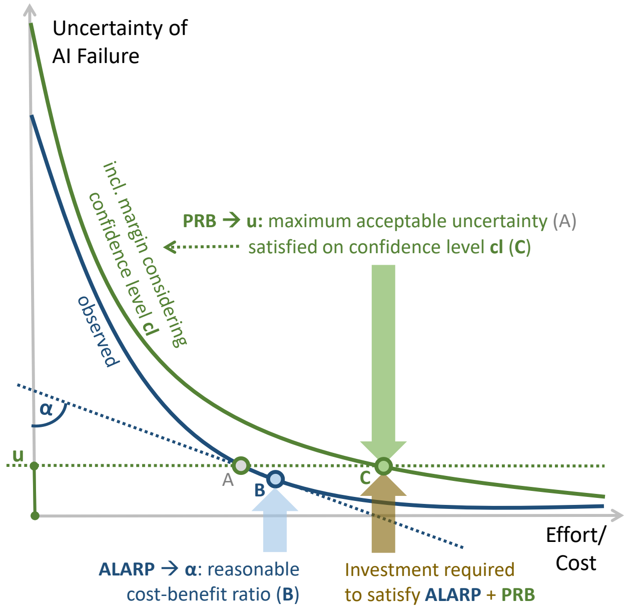

Uncertainty of AI Failure
incl. margin considering confidence level cl
observed
PRB → u: maximum acceptable uncertainty (A)
satisfied on confidence level cl (C)
α
u
A
B
C
Effort/ Cost
ALARP → α: reasonable cost-benefit ratio (B)
Investment required to satisfy ALARP + PRB

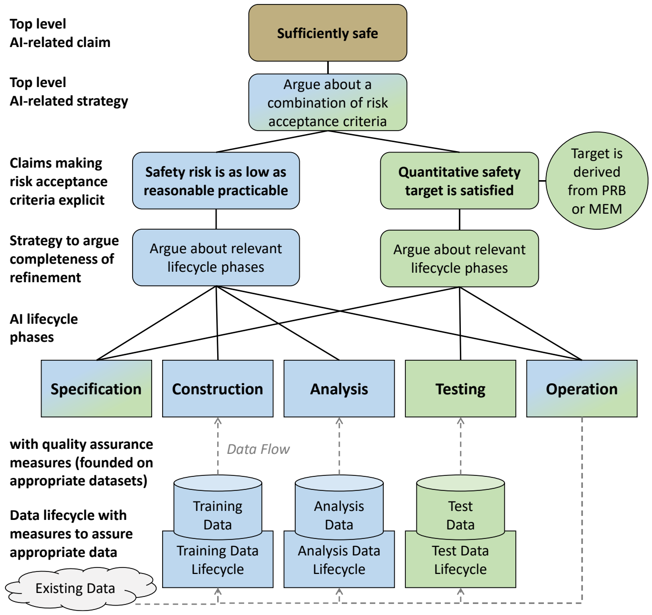
Figure 2: Implications of considering two risk acceptance criteria

4.3 Arguing considering the AI lifecycle

As illustrated above, it seems reasonable to argue two separate risk acceptance criteria. It is also advisable to argue each criterion independently. Important for the argumentation, especially for the argumentation of ALARP, are strategies that assure that the refinement of the claims into sub-claims is complete. An accepted way, which we also consider as most promising, is to use a lifecycle model to argue completeness and localize safety measures.

The lifecycle model for AI components presented in Fig. 3 builds on existing work, in particular on the work of Ashmore et al. [2019] and Gauerhof et al. [2020]. We adapted their proposals. The objective was to achieve an even clearer separation and better assignability of datasets, objectives and corresponding safety measures to the individual phases. In addition, we tried to keep the phases sufficiently generic to be applicable for the various development processes in data science projects that we are aware of.

We distinguish between *specification*, *construction*, *analysis*, *testing*, and *operation*. The proposed lifecycle model explicitly does not include a 'data' phase. Subsuming data-related activities in single phase neither matches reality nor gives weight to the topic of data, which is at the core of any AI lifecycle. Especially since different data with different qualities are consumed in different phases, we modeled individual data lifecycles as parallel streams that provide the foundation for the evidences created in the AI lifecycle.



Top level AI-related claim
Sufficiently safe
Top level AI-related strategy
Argue about a combination of risk acceptance criteria
Claims making risk acceptance criteria explicit
Safety risk is as low as reasonable practicable
Quantitative safety target is satisfied
Target is derived from PRB or MEM
Strategy to argue completeness of refinement
Argue about relevant lifecycle phases
Argue about relevant lifecycle phases
AI lifecycle phases
Specification
Construction
Analysis
Testing
Operation
with quality assurance measures (founded on appropriate datasets)
Data Flow
Data lifecycle with measures to assure appropriate data
Training Data
Analysis Data
Test Data
Training Data Lifecycle
Analysis Data Lifecycle
Test Data Lifecycle
Existing Data

Figure 3: AI lifecycle phases with mapping to risk criteria

The proposed separation also results in the fact that certain phases exclusively contribute either to argue ALARP or the target-based risk acceptance criterion, as we will show.

Specification considers, among other things, the definition of the AI task, the target application scope [Kläs and Vollmer, 2018], which is comparable to the operation design domain in automotive, and safety-relevant as well as other quality requirements including system constraints like computational resources. Although the AI specification has some specialties, activities are largely AI-independent. Nevertheless, it is a key phase for both types of risk acceptance criteria. A sufficiently complete and correct specification is a prerequisite to assuring that the safety risk will be as low as reasonable practicable by proving guardrails for the subsequent phases, but it also constitutes the AI-specific *safety target* and the scope in which this target has to be achieved. For example, *the AI must not falsely detect a signal for the right of way that was not actually given in more than one of N cases in its previously defined target application scope*.

Construction is an AI-specific phase. Its objective is to build a model from a training dataset that is able to fulfill the AI task in the target application scope considering the requirements and constraints defined in the specification.

During construction, many design decisions have to be made, e.g., on the kind of model and its hyperparameters including topology, learning algorithm, stop criteria, etc.

Many of these decisions are based on trial and error, taking into account experience, so construction is a highly iterative process in a close feedback loop with the analysis phase.

We will not be able to show during construction that we achieved a certain quantitative target, since we focus on fitting but not soundly testing the model. Thus, this phase plays no role in arguing regarding the target-based risk acceptance criteria. However, considering quality assurance measures during construction is important to argue ALARP. The applied quality assurance measures should be guided by the safety target, but also by other quality requirements and constraints identified during the specification. Commonly, it is not possible to define fixed success criteria for the different quality measures. For example, in most cases, it would not be reasonable to enforce a specific type of model or topology, or request a maximum batch size *m* and run at least *e* epochs. Instead, we propose analyzing and monitoring the efficiency of the measures carried out and stopping in accordance with ALARP if a reasonable saturation is achieved. For example, if performing a random search on appropriate hyperparameter values, the search shall continue as long as the model shows reasonable improvements.

Analysis is also an AI-specific phase that is performed in a close feedback loop with construction to provide guardrails for improving construction and indicating the achievement of saturation for constructive quality assurance measures. Analysis comprises besides means for explainability also “testing” the model on validation data to estimate and monitor the model performance with respect to the safety target. However, although techniques are applied that are similar to the techniques applied in the testing phase, the analysis phase differs from the testing phase in that objective is to gather insides to further improve the AI model rather than provide evidence for the achievement of the specified safety target. Therefore, the quality assurance measures in the analysis phase help to argue ALARP but do not contribute to arguing regarding the target-based risk acceptance criteria. In analogy to the construction phase, it is difficult to define a priori targets for most quality measures in the analysis phase. Rather, their effect and thus their potential contribution to the safety target must be monitored and continuously evaluated.

Testing is also commonly considered to be AI-specific. Unlike analysis, the objective of the testing phase is to generate evidences on the achievement of the quantitative safety target. In providing these evidences, testing depends on the specification, including the definition of the AI task and the target application scope. Moreover, it relies on specific qualities of the test data that are not so relevant, for example, for training data, such as that the data fulfills some representativeness criteria and that it was not used previously during construction or analysis. Since a test dataset can always provide only a sample of all possible cases in the target application scope, we need to underpin the evidence on satisfying the safety target with some statistical confidence (cf. Fig. 2) [Kläs and Sembach, 2019]. The confidence level *cl*, which is independent of the target, may be set based on criticality or requested integrity. For example, we might request that the probability that we falsely confirm our target '*The AI must not falsely detect a signal for the right of way that was not actually given in more than one of N cases in its previously defined target application scope*.' is less than 1-*cl* = 0.0001.

Moreover, it is important to understand that quality assurance measures in the testing phase are not applied to further improve the AI solution and thus does not provide evidences to argue ALARP. Instead, they help argue that we are confident that we have met the quantitative safety target.

Operation in the sense considered here comprises deployment, usage, maintenance, and retirement. Although most aspects are not AI-specific, some are and need to be addressed with appropriate safety measures. On the one hand, measures for assuring ALARP include the collection of relevant information during operation to further improve the AI solution as part of maintenance. Moreover, situations have to be detected in which the AI solution can only provide outcomes with high uncertainty, in order to allow for appropriate countermeasures to be taken on the system level to improve the overall safety. Such situations may include settings where lighting conditions make falsely detecting a signal for the right of way much more likely. On the other hand, evidence on satisfying the safety target obtained from testing strongly relies on assumptions regarding the target application scope; if the AI solution is applied in a different setting or relevant characteristics of the application change, this evidence is no longer valid. Therefore, safety measures have to be taken during operation to detect such deviation between the target application scope and the actual application scope.

5 Discussion

We proposed a strategy for arguing the safety of an AI-based safety function combining two risk acceptance criteria. The structure can help to come up with a sound argument but there are ways of how one could attack this argument. A possible attack on the ALARP argument is that the body of knowledge concerning the effectiveness and the best combinations of measures is not mature enough. A possible attack on the quantitative claim based on PRB or MEM is that there is not enough practical experience and empirical evidence. A possible response to this attack is to collect data during operation and to use market monitoring to strengthen the argument. This approach is described already by the Safety Performance Indicator [Koopman and Wagner, 2019] or GQM$^+$Strategies [Basili et al., 2010] but it needs to be tailored to the focused argument for AI. By evaluating the reasoning with data, a mature body of knowledge can be developed over time and reflected in safety standards for AI.

Considering standardization, we see three options for using ACs. The first is to demand in a safety standard the development of an AC for the considered product. The second is to describe in product- or domain-specific safety standards a generic AC that shall be instantiated. The third is to develop a product- or domain-specific AC and use this AC to develop a checklist-based safety standard where safety measures are chosen depending on the specific criticality/integrity level.

Considering certification, we see two main aspects. The first is that the AC needs to comply with the standard describing what the AC should look like. The second and more important aspect is that the AC itself needs to be sound, so that it can be accepted by the certification body. The challenge here is that the review of the AC becomes easily more elaborative than a checklist-based approach, meaning the certification body needs much greater expertise. Furthermore, the certification body can no longer give up responsibility for the safety of the system by saying that it is only responsible for compliance with standards but not for system safety. However, this aspect is not specific for AI and is generally true for the certification of complex systems by means of ACs.

6 Conclusion

We conclude that ACs have the potential to justify the usage of AI in safety-critical systems. A prerequisite is, however, that they argue that risks are as low as reasonably practicable (ALARP) and that a reasonable target based on a quantitative risk acceptance criterion has been chosen and is fulfilled. We presented the first approach for explicitly augmenting the achievement of these complementary objectives for AI.

We also see the potential of the proposed structure for traditional software as it would enforce claims about the effectiveness of safety measures. It would put into question whether one is really following the ALARP principle when choosing safety measures according to recommendations given by safety standards. It would also raise the question of how effective software safety measures are and call for empirical evidences about their effectiveness.

Last but not least, we advocate that the concept of ACs from the safety community should be carried over to the AI community. In particular, researchers with a background in empirical studies and data quality need to be involved in the development and review of AI-related ACs.

Acknowledgments

Parts of this work have been funded by the Observatory for Artificial Intelligence in Work and Society (KIO) of the Denkfabrik Digitale Arbeitsgesellschaft in the project "KI Testing & Auditing". We would also like to thank Sonnhild Namingha for an initial review of this paper.

References

[Adelard LLP, 2021] Adelard LLP. *CAE FRAMEWORK*, 2021, https://claimsargumentsevidence.org/. Accessed 10 May 2021.

[Adler et al., 2019] R. Adler, M. N. Akram, P. Bauer, p: Feth, P. Gerber, A. Jedlitschka, L. Jöckel, M. Kläs, and D. Schneider. *Hardening of Artificial Neural Networks for Use in Safety-Critical Applications - A Mapping Study*, 2019. https://arxiv.org/abs/1909.03036.

[Ashmore et al., 2019] R. Ashmore, R. Calinescu, and C. Paterson. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges, *ACM Computing Surveys*, 2019.

[Basili et al., 2010] V. R. Basili, et al. Linking Software Development and Business Strategy Through Measurement. *Computer*. 43(4):57–65, 2010.

[BSI, 2021] BSI, Fraunhofer HHI, Verband der TÜV. *Towards Auditable AI Systems*, 2021.

[DIN, 2017] *DIN EN ISO 3691-1:2017 – Industrial trucks - Safety requirements and verification*, 2017.

[European Commission, 2021] European Commission. *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, 2021. https://ec.europa.eu/newsroom/dae/redirection/item/709090.

[Felderer and Ramler, 2021] M. Felderer, and R. Ramler. Quality Assurance for AI-based Systems: Overview and Challenges. In *Software Quality: Future Perspectives on Software Engineering Quality*. 2021.

[Gauerhof et al., 2020] L. Gauerhof, R. Hawkins, C. Picardi, C. Paterson, and I. Habli. Assuring the Safety of Machine Learning for Pedestrian Detection at Crossings. In *Proc. of SAFECOMP 2020.* Springer, pp. 197–212, 2020.

[Gauerhof et al., 2018] L. Gauerhof, P. Munk, and S. Burton. Structuring Validation Targets of a Machine Learning Function Applied to Automated Driving. In *Proc. of SAFECOMP 2018*, pp. 45–58, 2018.

[Hauer et al., 2021] Hauer, M., Adler, R., and Zweig, K. Assuring Fairness of Algorithmic Decision Making (ITEQS 2021). In *Proc. of Int. Conf. on Software Testing*, 2021.

[Hawkins et al., 2021] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli. *Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS), 2021*.

[IEC 2010] IEC 61508-5:2010 – Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems, 2010.

[INI, 2019] Institut für Neuroinformatik. *German Traffic Sign Benchmarks*, 2019. https://benchmark.ini.rub.de/gtsrb_results.html. Accessed 10 May 2021.

[ISO/TR, 2020] ISO/TR 4804:2020 – Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation, 2020.

[ISO/IEC, 2021] ISO/IEC AWI TR 5469 – Artificial intelligence — Functional safety and AI systems, 2021.

[ISO/IEC/IEEE, 2019] ISO/IEC/IEEE 15026-1:2019 – Systems and software engineering - Systems and software assurance - Part 1: Concepts and vocabulary, 2019.

[Kläs and Sembach, 2019] M. Kläs, and L. Sembach. Uncertainty Wrappers for Data-Driven Models. In *Proc. of SAFECOMP 2019.* Springer, pp. 358–364, 2019.

[Kläs and Vollmer, 2018] M. Kläs, and A.M. Vollmer. Uncertainty in Machine Learning Applications: A Practice-Driven Classification of Uncertainty. In Proc. of *SAFECOMP 2019, 2019.*

[Mayr et al., 2012] A. Mayr, R. Plösch, M. Kläs, C. Lampasona, and M. Saft. A Comprehensive Code-Based Quality Model for Embedded Systems: Systematic Development and Validation by Industrial Projects. In *ISSRE 2012*, pp. 281-290, 2012.

[OMG, 2020] Object Management Group. *About the Structured Assurance Case Metamodel Specification Version 2.1*, 2020.

[Koopman and Wagner, 2019] P. Koopman, and M. Wagner. Positive Trust Balance for Self-driving Car Deployment. In *Proc. of SAFECOMP 2020 Workshops*, pp. 351–357, 2019.

[Picardi et al., 2019] C. Picardi, R. Hawkins, C. Paterson, and I. Habli. A Pattern for Arguing the Assurance of Machine Learning in Medical Diagnosis Systems. In *Proc. of SAFECOMP 2019*, pp. 165–179, 2019.

[Picardi et al., 2020] C. Picardi, C. Paterson, R. Hawkins, R. Calinescu, and I. Habli. Assurance Argument Patterns and Processes for Machine Learning in Safety-Related Systems. In *Proc. of SafeAI 2020*, pp. 23-30, 2020.

[Pietsch, 2021] B. Pietsch. 2 Killed in Driverless Tesla Car Crash, Officials Say. *The New York Times, 2021.*

[SCSC, 2018] Safety-Critical Systems Club. *GSN Community Standard Version 2 Draft 1*, 2018.

[Salay and Czarnecki, 2018] R. Salay, and K. Czarnecki. *Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262, 2018.* https://arxiv.org/abs/1808.01614.

[Sämann et al., 2020] T. Sämann, P. Schlicht, and F. Hüger. *Strategy to Increase the Safety of a DNN-based Perception for HAD Systems, 2020.* https://arxiv.org/abs/2002.08935.

[Schwalbe et al., 2020] G. Schwalbe, et al. Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications. In *Proc. of SAFECOMP 2020*, pp. 383–394, 2020.

[Schwalbe and Schels, 2020] G. Schwalbe, and M. Schels. A Survey on Methods for the Safety Assurance of Machine Learning Based Systems. In: *Proc. of European Congress on Embedded Real Time Software and Systems, 2020.*

[Schwalbe and Schels, 2019] G. Schwalbe, and M. Schels. Strategies for Safety Goal Decomposition for Neural Networks. In *Proc. of ACM Computer Science in Cars Symposium, 2020*.

[Siebert et al., 2021] J. Siebert, L. Joeckel, J. Heidrich, A. Trendowicz, K. Nakamichi, K. Ohashi, I. Namba., R. Yamamoto, and M. Aoyama. Construction of a Quality Model for Machine Learning Systems, *Software Quality Journal*. Special Issue Information Systems Quality, 2021.

[UL, 2021] Underwriters Laboratories. *Presenting the Standard for Safety for the Evaluation of Autonomous Vehicles and Other Products*. https://ul.org/UL4600. Accessed 10 May 2021.

[VDE, 2020] VDE-AR-E 2842-61-1:2020-07 – Development and trustworthiness of autonomous/cognitive systems, 2020.

[Wakabayashi, 2018] D. Wakabayashi. Self-Driving Uber Car Kills Pedestrian in Arizona. *The New York Times*, 2018.

[Willers et al., 2020] O. Willers, S. Sudholt, S. Raafatnia, and S. Abrecht. Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks. In *Proc. of SAFECOMP 2020*, pp. 336–350, 2020.

[Wozniak *et al.*, 2020] E. Wozniak, C. Cârlan, E. Acar-Celik, and H. Putzer. A Safety Case Pattern for Systems with Machine Learning Components. In *Proc. of SAFECOMP 2020.* Springer, pp. 370-382, 2020.

[Zenzic, 2020] Zenzic-UK Ltd. *Zenzic-Safety-Framework-Report-2.0-final*, 2020. https://zenzic.io/reports-and-resources/safety-case-framework/. Accessed 10 May 2021.