Fraunhofer IESE

Fraunhofer Institute for Experimental Software Engineering IESE

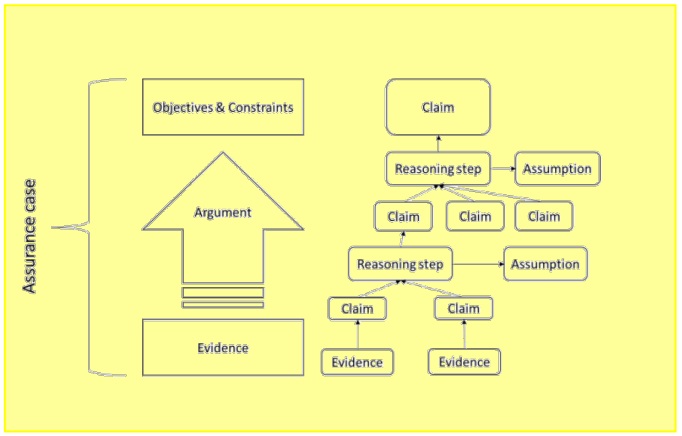
A Framework to Argue Quantitative Safety Targets in Assurance Cases for AI/ML Components Combining Design and Runtime Safety Measures

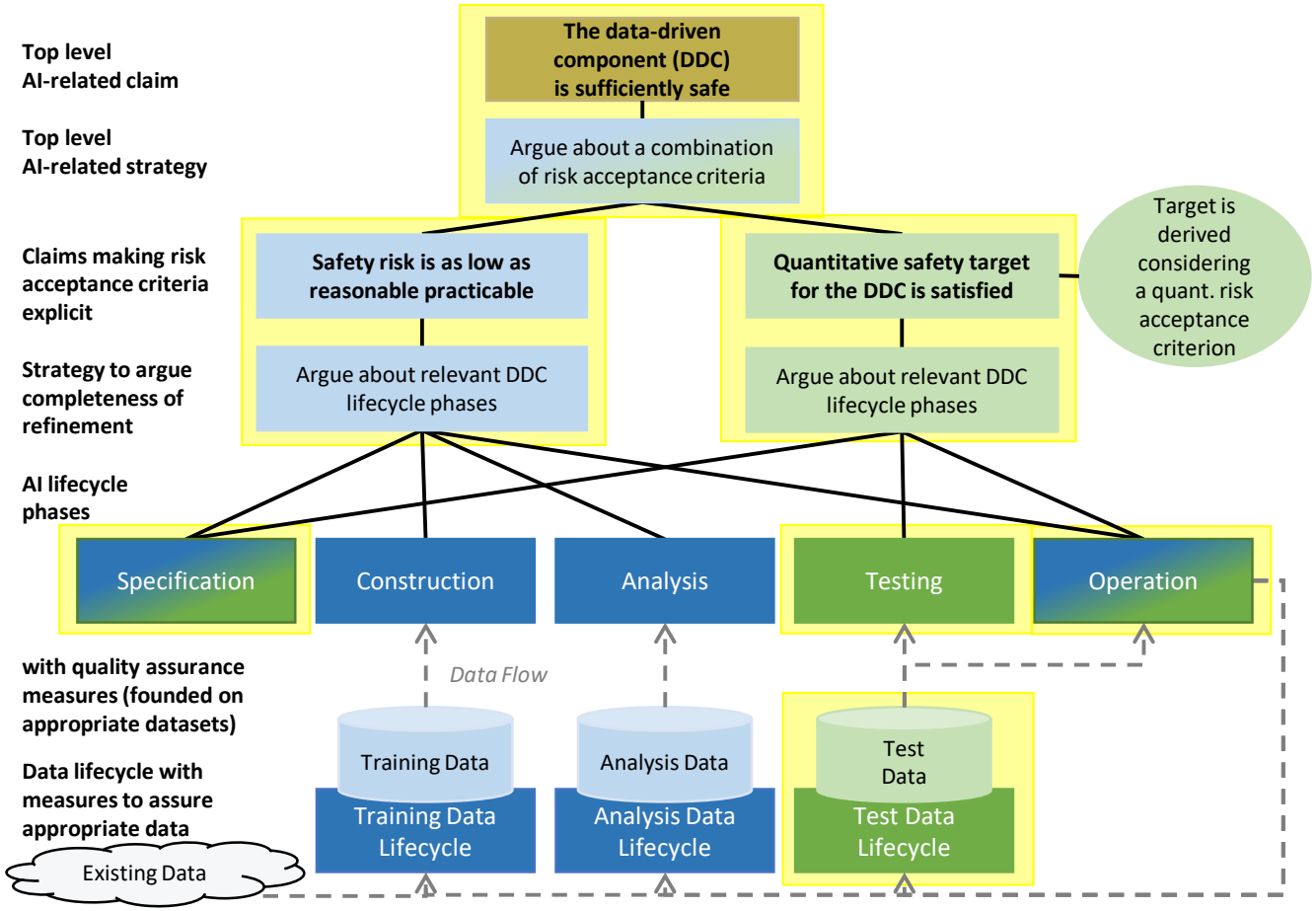
SafeAI 2022 : The AAAI's Workshop on Artificial Intelligence Safety, March 1, 2022

Michael Kläs, Lisa Jöckel, Rasmus Adler, Jan Reich
{michael.klaes, lisa.joeckel, rasmus.adler, jan.reich}@iese.fraunhofer.de

Assurance cases structured by complementary risk acceptance criteria

are a means to argue safety of an AI-enabled system

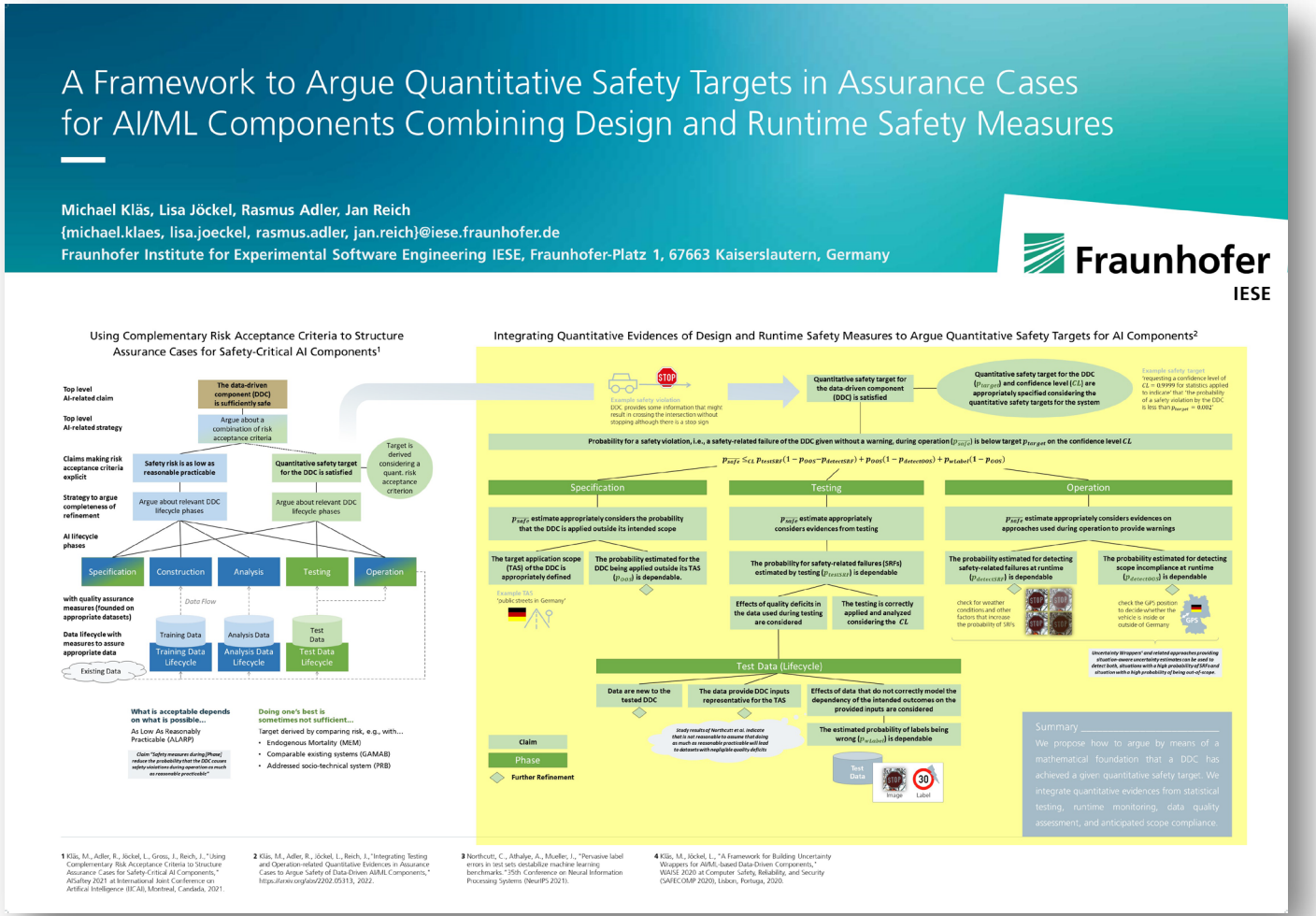




Kläs, M., Adler, R., Jöckel, L., Gross, J., Reich, J., "Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components," AISaftey 2021 at International Joint Conference on Artifical Intelligence (IJCAI)

An argument structure to mathematically integrate quantitative evidences

considering evidences on data quality, design and runtime safety measures



A Framework to Argue Quantitative Safety Targets in Assurance Cases for AI/ML Components Combining Design and Runtime Safety Measures

Michael Kläs, Lisa Jöckel, Rasmus Adler, Jan Reich
{michael.klaes, lisa.joeckel, rasmus.adler, jan.reich}@iese.fraunhofer.de
Fraunhofer Institute for Experimental Software Engineering IESE, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany

Fraunhofer IESE

Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components[1]

Integrating Quantitative Evidences of Design and Runtime Safety Measures to Argue Quantitative Safety Targets for AI Components[2]

Integrating Testing and Operation-related Quantitative Evidences in Assurance Cases to Argue Safety of Data-Driven AI/ML Components

Michael Kläs, Lisa Jöckel, Rasmus Adler, Jan Reich
Fraunhofer Institute for Experimental Software Engineering IESE,
Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
{michael.klaes, lisa.joeckel, rasmus.adler, jan.reich}@iese.fraunhofer.de

Abstract

In the future, AI will increasingly find its way into systems that can potentially cause physical harm to humans. For such safety-critical systems, it must be demonstrated that their residual risk does not exceed what is acceptable. This includes, in particular, the AI components that are part of such systems' safety-related functions. Assurance cases are an intensively discussed option today for specifying a sound and comprehensive safety argument to demonstrate a system's safety. In previous work, it has been suggested to argue safety for AI components by structuring assurance cases based on two complementary risk acceptance criteria. One of these criteria is used to derive quantitative targets regarding the AI. The argumentation structures commonly proposed to show the achievement of such quantitative targets, however, focus on failure rates from statistical testing. Further important aspects are only considered in a qualitative manner – if at all. In contrast, this paper proposes a more holistic argumentation structure for having achieved the target, namely a structure that integrates test results with runtime aspects and the impact of scope compliance and test data quality in a quantitative manner. We elaborate different argumentation options, present the underlying mathematical considerations, and discuss resulting implications for their practical application. Using the proposed argumentation structure might not only increase the integrity of assurance cases but may also allow claims on quantitative targets that would not be justifiable otherwise.

1. Motivation

Components based on machine learning (ML) or artificial intelligence (AI) are increasingly necessary for many innovative, especially autonomous, systems, like self-driving vehicles in complex environments. They provide features that could not be realized (with competitive quality) using traditional software. It is almost unavoidable that these data-driven components (DDC) become safety-critical. Safety architectures can lower the criticality of a DDC but it is often hardly possible to get a DDC completely out of the safety-critical path. Currently, standards are missing for AI-enabled systems, but we need to find ways to assure that the AI-enabled system is sufficiently safe in its context.

Assurance refers to "grounds for justified confidence that a claim has been or will be achieved". Using assurance cases – as an established approach in safety engineering – also for this purpose is a heavily discussed option. An assurance case is defined as a "reasoned, auditable artefact created that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s)" [ISO/IEC/IEEE, 2019]. Assurance cases are a flexible means for dealing with standardization gaps and provide a structured way of arguing quality requirements of a (data-driven) software component. Using assurance cases for arguing safety if a DDC is in the safety-critical path is considered in current safety standardization (e.g., NWIP ISO/PAS 8800, VDE-AR-E 2842-61), research projects (e.g., KI Absicherung, EXAMAI), and communities (e.g., Safety-critical systems club).

In our previous work, we proposed a possible overall structuring of the argumentation of an assurance case based on complementary risk acceptance criteria [Kläs et al., 2021]. This means that there are two separate lines of assurance case argumentation, one following the "as low as reasonably practicable (ALARP)" risk acceptance criteria and the second one using a quantitative target, e.g., a sufficiently low probability for such AI outcomes that may affect safety.

Existing assurance case structures for arguing the quantitative targets mainly focus on evidences provided by statistical testing. Other important aspects contributing to achieving the quantitative target are not integrated in its argumentation but at most considered exclusively in qualitative way.

We see two problems with this: (a) On the one hand, ignoring such aspects as part of a quantitative argumentation

Michael Kläs, Lisa Jöckel, Rasmus Adler, Jan Reich, "Integrating Testing and Operation-related Quantitative Evidences in Assurance Cases to Argue Safety of Data-Driven AI/ML Components", https://arxiv.org/abs/2202.05313

Fraunhofer IESE