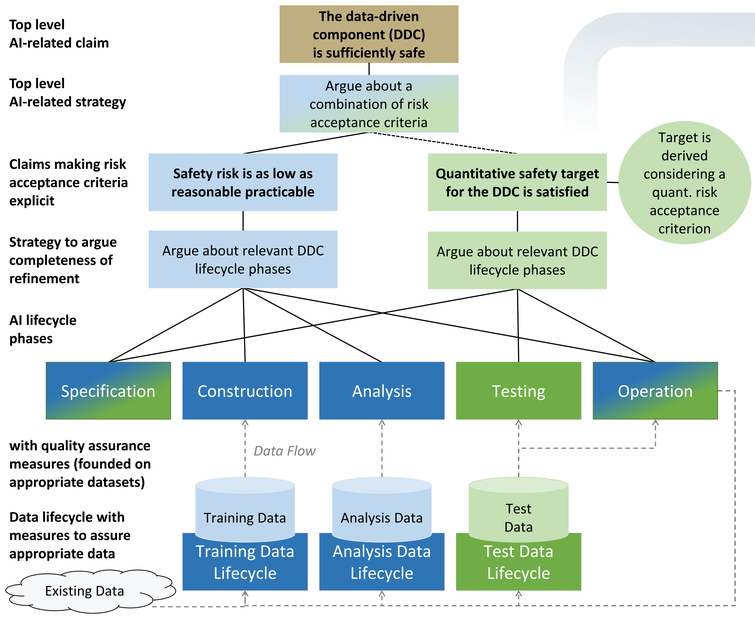
An Assurance Case Pattern to Argue Quantitative Safety Targets for AI Components Considering Their Complete Life-Cycle

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]



What is acceptable depends on what is possible...

As Low As Reasonably Practicable (ALARP)

Claim "Safety measures during [Phase] reduce the probability that the DDC causes safety violations during operation as much as reasonably practicable"

Doing one's best is sometimes not sufficient...

Target derived by comparing risk, e.g., with...

- Endogenous Mortality (MEM)
- Comparable existing systems (GAMAB)
- Addressed socio-technical system (PRB)

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

Example safety violation
DDC provides some information that might result in crossing the intersection without stopping although there is a stop sign

Quantitative safety target for the data-driven component (DDC) is satisfied

Quantitative safety target for the DDC (p_{target}) and confidence level (CL) are appropriately specified considering the quantitative safety targets for the system

Example safety target
'requesting a confidence level of $CL = 0.9999$ for statistics applied to indicate' that 'the probability of a safety violation by the DDC is less than $p_{target} = 0.002$'

Probability for a safety violation, i.e., a safety-related failure of the DDC given without a warning, during operation ($p_{\overline{safe}}$) is below target p_{target} on the confidence level CL

$$p_{\overline{safe}} \leq_{CL} p_{testSRF}(1 - p_{OOS} - p_{detectSRF}) + p_{OOS}(1 - p_{detectOOS}) + p_{wLabel}(1 - p_{OOS})$$

Specification

$p_{\overline{safe}}$ estimate appropriately considers the probability that the DDC is applied outside its intended scope

The target application scope (TAS) of the DDC is appropriately defined

Example TAS
'public streets in Germany'

The probability estimated for the DDC being applied outside its TAS (p_{OOS}) is dependable.

Testing

$p_{\overline{safe}}$ estimate appropriately considers evidences from testing

The probability for safety-related failures (SRFs) estimated by testing ($p_{testSRF}$) is dependable

Effects of quality deficits in the data used during testing are considered

The testing is correctly applied and analyzed considering the CL

Operation

$p_{\overline{safe}}$ estimate appropriately considers evidences on approaches used during operation to provide warnings

The probability estimated for detecting safety-related failures at runtime ($p_{detectSRF}$) is dependable

check for weather conditions and other factors that increase the probability of SRFs

The probability estimated for detecting scope incompliance at runtime ($p_{detectOOS}$) is dependable

check the GPS position to decide whether the vehicle is inside or outside of Germany

Uncertainty Wrappers[4] and related approaches providing situation-aware uncertainty estimates can be used to detect both, situations with a high probability of SRFs and situation with a high probability of being out-of-scope.

Test Data (Lifecycle)

Data are new to the tested DDC

The data provide DDC inputs representative for the TAS

Study results of Northcutt et al. indicate that is not reasonable to assume that doing as much as reasonable practicable will lead to datasets with negligible quality deficits

Effects of data that do not correctly model the dependency of the intended outcomes on the provided inputs are considered

The estimated probability of labels being wrong (p_{wLabel}) is dependable

Claim
Phase
Further Refinement

Summary

We propose how to argue by means of a mathematical foundation that a DDC has achieved a given quantitative safety target. We integrate quantitative evidences from statistical testing, runtime monitoring, data quality assessment, and anticipated scope compliance.

1 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), Montreal, Candada, 2021.

2 Kläs, M., Adler, R., Jöckel, L., Reich, J., "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, 2022.

3 Northcutt, C., Athalye, A., Mueller, J., "Pervasive label errors in test sets destabilize machine learning benchmarks."35th Conference on Neural Information Processing Systems (NeurIPS 2021).

4 Kläs, M., Jöckel, L., "A Framework for Building Uncertainty Wrappers for AI/ML-based Data-Driven Components," WAISE 2020 at Computer Safety, Reliability, and Security (SAFECOMP 2020), Lisbon, Portuga, 2020.

Please check for details
https://arxiv.org/abs/2202.05313