USING COMPLEMENTARY RISK ACCEPTANCE CRITERIA TO STRUCTURE ASSURANCE CASES FOR SAFETY-CRITICAL AI COMPONENTS

Michael Kläs, Rasmus Adler, Lisa Jöckel, Janek Groß, Jan Reich

AI Safety 2021, 20 August 2021, Virtual
Assurance Cases (ACs) can support the upcoming need to certify AI

- AI becomes safety-related i.e., its failure significantly increase risk for people or environment

- Upcoming regulations e.g., the European AI Act, require the certification of high-risk AI

- Currently no standards exist that can be easily applied to AI

- Assurance cases as an established approach in safety engineering may provide an alternative

https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

https://arxiv.org/abs/1709.02435
ACs can be described in a tree structure

(Top-level) Claim

Sufficiently safe

Argument strategy

Sub claim...

Sub claim

Argument strategy

Sub claim

Sub claim

Evidence


We see some common weaknesses in the structuring of AI-related AC

How to justify that we are complete in the considered sources?

How to justify that the probability of hazards caused by this source is sufficiently reduced?

What is the target value for a given indicator such as neuron coverage?

Algorithm performance issues
- Argue about prevention or mitigation approaches

Incomprehensible behavior
- Limited practical experience and empirical data on cause-effect relationships

Sufficiently safe
- Argue about a low probability of hazards by considering potential sources / insufficiencies

AI behavior range is fully covered as measured by neuron coverage (NC)

NC = 100%

...
We consider complementary risk acceptance criteria for AI to be essential

Safety as “freedom from unacceptable risk” (of physical injury)

Safety risk is as low as reasonable practicable

Quantitative safety target is satisfied

Target is derived from PRB, MEM,…

Doing one’s best is sometimes not sufficient…
Target derived by comparing risk with…
• Endogenous Mortality (MEM)
• Comparable existing systems (GAMAB)
• Addressed socio-technical system (PRB)

What is acceptable depends on what is possible…
As Low As Reasonably Practicable (ALARP)

Sufficiently safe

Argue about complementary risk acceptance criteria

[IEC 61508]
Risk acceptance criteria can be mapped on a generic AI lifecycle

- **Sufficiently safe**
- Argue about complementary risk acceptance criteria
- **Safety risk is as low as reasonable practicable**
- **Quantitative safety target is satisfied**

Strategy to argue completeness of refinement

**AI lifecycle phases**
- Specification
- Construction
- Analysis
- Testing
- Operation

**Data lifecycle with measures to assure appropriate data**
- Training Data Lifecycle
- Analysis Data Lifecycle
- Test Data Lifecycle
Risk acceptance criteria can be mapped on a generic AI lifecycle

- **Sufficiently safe**
  - Argue about complementary risk acceptance criteria
    - Safety risk is as low as reasonable practicable
      - Argue about relevant lifecycle phases
    - Quantitative safety target is satisfied

**Strategy to argue completeness of refinement**

- **AI lifecycle phases**
  - Specification
  - Construction
  - Analysis
  - Testing
  - Operation

**Data lifecycle with measures to assure appropriate data**

- Training Data Lifecycle
- Analysis Data Lifecycle
- Test Data Lifecycle

**Claim**

"Safety measures during [Phase] reduce the probability that the AI causes safety-related failures during operation as much as reasonable practicable"
Risk acceptance criteria can be mapped on a generic AI lifecycle

- **Sufficiently safe**
  - Argue about complementary risk acceptance criteria
    - Safety risk is as low as reasonable practicable
    - Quantitative safety target is satisfied
      - Argue about relevant lifecycle phases

**AI lifecycle phases**
- Specification
- Construction
- Analysis
- Testing
- Operation

**Data lifecycle with measures to assure appropriate data**
- Training Data Lifecycle
- Analysis Data Lifecycle
- Test Data Lifecycle

Strategy to argue completeness of refinement

Thank you for your interest

Our proposal in a nutshell
- Split argumentation using complementary risk acceptance criteria
- Argue ALARP for the each relevant phase of the AI lifecycle
- Use saturation as a cost-benefit-based stop criteria for safety measures
- Show achievement of the quantitative safety target during testing
- Monitor assumptions regarding the application scope during operation

Interested in more details? Check our paper or contact us!
- How complementary risk acceptance criteria influence cost and risk
- Possible scenarios on how ACs can support AI certification

Dr. Michael Kläs
michael.klaes@iese.fraunhofer.de
Phone: +49 631 / 6800-2110
Fraunhofer IESE
Kaiserslautern, Germany