Safety Assurance & AI in the Automotive Domain
- AI Standards
- Example: AI-Based SoC estimation for EVs

Fredrik Warg <fredrik.warg@ri.se>

Martin Skoglund, Aria Mirzai, Anders Thorsén, Karl Lundgren, Peter Folkesson, Bastian Havers-Zulka

Context

Trustworthy AI[1]

3 components
- Lawful
- Ethical
- Robust

4 ethical principles
- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

7 key requirements
- Accountability
- Privacy and data governance
- Transparency
- Technical robustness and safety
- Human agency and oversight
- Societal and environmental wellbeing
- Diversity, non-discrimination, fairness

AI Act[2]

Unacceptable risk
(Prohibited)
E.g. social scoring, manipulative, deceptive

High risk
(Regulated)
E.g. safety components, biometrics, critical infra.

Limited risk
(Transparency obligations)
E.g. chatbots and generative AI content

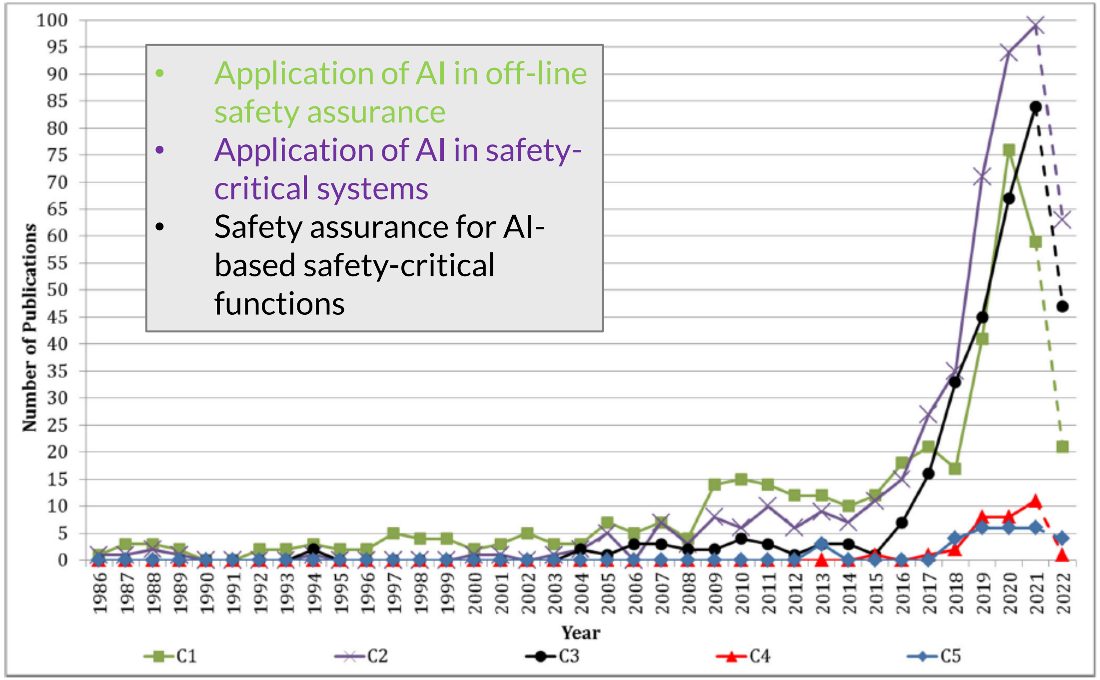
Minimal risk
(Unregulated)
E.g. simpler game AI, simple photo filters

[1] *https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai*

[2] *https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng*

AI in safety-critical systems

AI in safety-critical systems



Source: A. V. Silva Neto et al.: Safety Assurance of AI-Based Systems : A Systematic Literature Review on the State of the Art and Guidelines for Future Work, 2022.

Test tool

- Test case generation
- Analysis of results

Component in deployed system

- Object detection
- Decision-making
- Decision support

Development tool

- Coding
- Architecture

Safety analysis

- Automated analysis
- Assessment tools

AI Standardization

[System safety]
Information technology – Artificial intelligence – Guidance on risk management

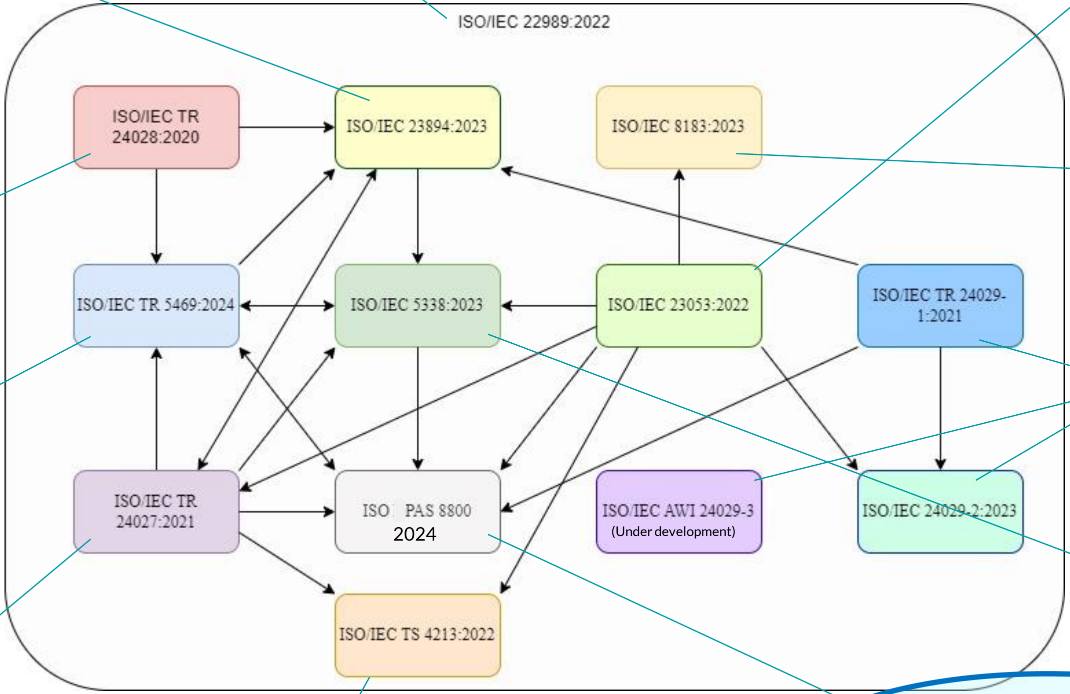
[Foundational]
Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

[Foundational]
Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

[Trustworthiness]
Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

[Functional safety]
Artificial intelligence – Functional safety and AI systems

[Trustworthiness]
Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making



[Life-cycle]
Information technology — Artificial intelligence — Data life cycle framework

[Trustworthiness]
Artificial Intelligence (AI) — Assessment of the robustness of neural networks

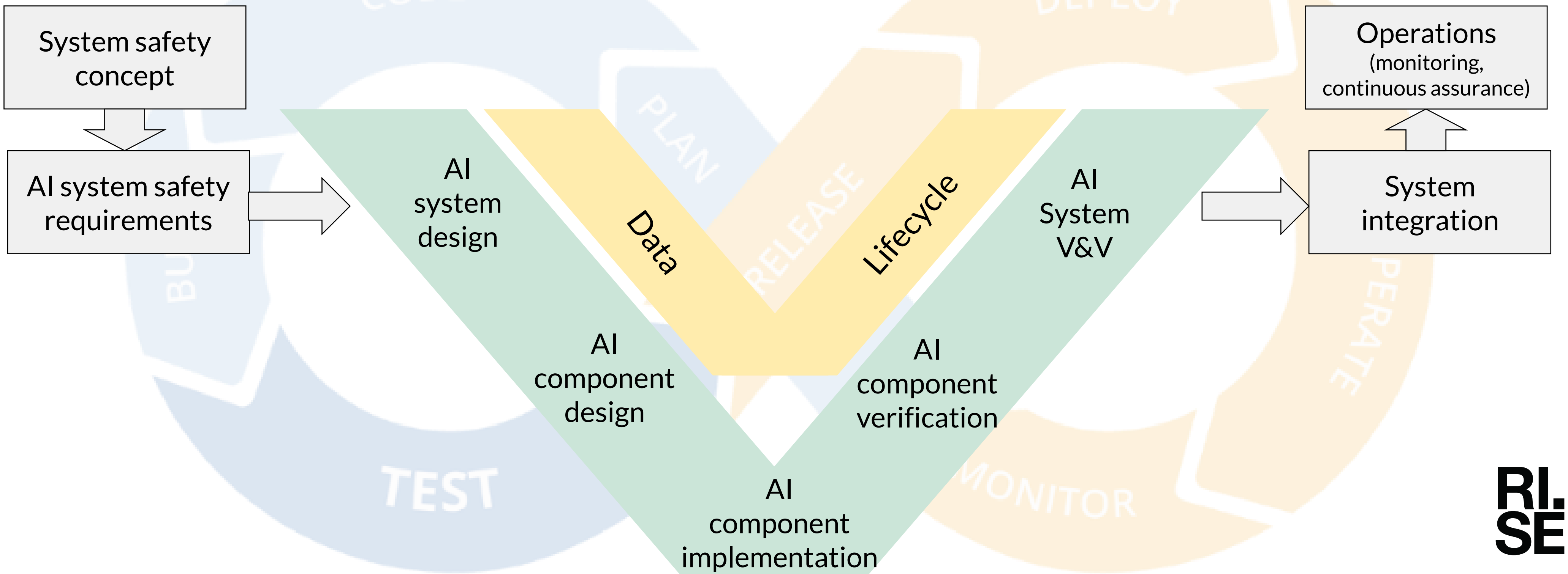
[Life-cycle]
Information technology — Artificial intelligence — AI system life cycle processes

[Quality]
Information technology — Artificial intelligence — Assessment of machine learning classification performance

[System safety]
Road vehicles — Safety and artificial intelligence

ISO/PAS 8800 Framework

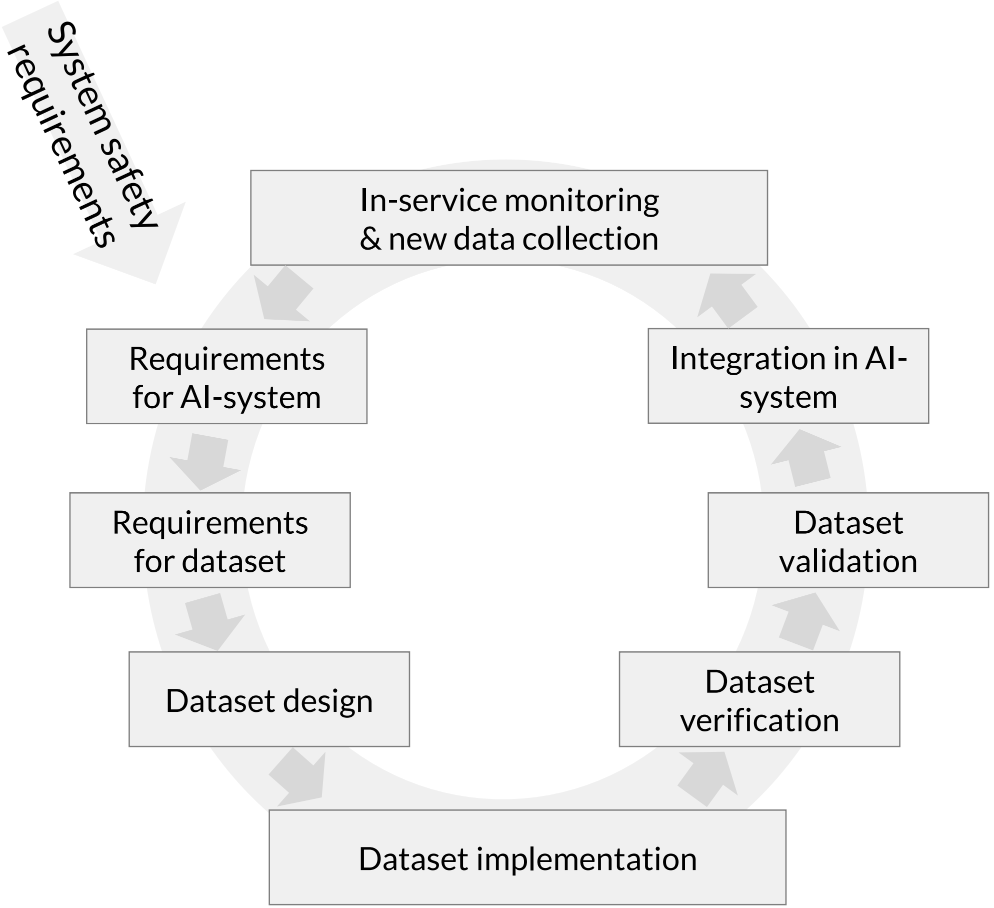
- AI used as part of safety related systems in road vehicles
- Not applicable to AI in software tools



RI. SE

Data Lifecycle

- Continuous lifecycle for post-deployment changes
  - Concept/data/semantic drift
  - Incidents/threats
- Data collection (pre- and post-deployment)
  - AI model training data
  - Test data
    - AI model test data
    - Scenario-based test data
- In-service monitoring and reporting (ISMR)
  - Metric/Incident reporting
  - Continuous risk assessment



RI.
SE

V&V Methods

- Choice of V&V methods based on multiple parameters
  - AI requirements
  - Test purpose
  - Model type
  - Model access
  - Learning paradigm
  - Type of task performed
- No fixed checklist in standards

Benchmarking
Standardized test suites. Performance is measured against annotated reference data or desired answers.

Robustness testing
Tests for robustness with respect to input data, e.g., simulating input noise.

Statistical testing
Evaluation of metrics defined within the AI safety requirements for the system

Edge cases
Testing values at the edge of the input space and unusual cases/combinations.

Sampling-based methods
Methods to guide testing to areas of the input space with higher error distribution

Explainability
Techniques to make the model's decisions (semi-)transparent. Can be used identify sources of unwanted behaviors.

Review/Expertise
Test cases constructed based on expert knowledge or based on model/data review.

Formal verification
Methods based on mathematical proofs to specify and verify properties.

Scenario-based tests
Stimulating model with collected data to evaluate real-world environment response

Gradient-based search
Use of knowledge of internal model parameters to guide generation of test cases

RI. SE

Case-study: AI in the Automotive Domain

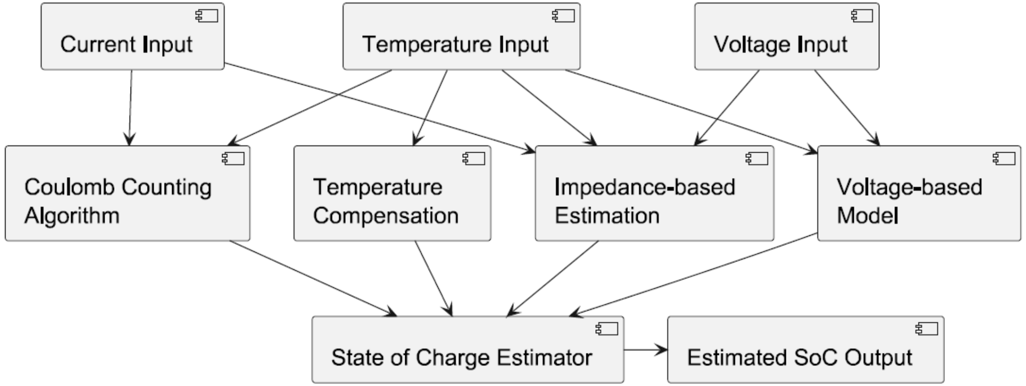
Case-study: State-of-Charge (SOC) Estimation

- SOC measures remaining charge
  - E.g., range information for an EV
- Critical functions
  - Prevent overcharging
  - Prevent deep discharging
- Worst case: Overcharging $\rightarrow$ heat generation $\rightarrow$ electrolyte decomposition $\rightarrow$ thermal runway $\rightarrow$ fire/toxic gases

From paper:
AI Safety Assurance in Electric Vehicles: A Case Study on AI-Driven SOC Estimation (EVS 38, June 2025) https://arxiv.org/abs/2509.03270
Martin Skoglund, Fredrik Warg, Aria Mirzai, Anders Thorsén, Karl Lundgren, Peter Folkesson, Bastian Havers-Zulka

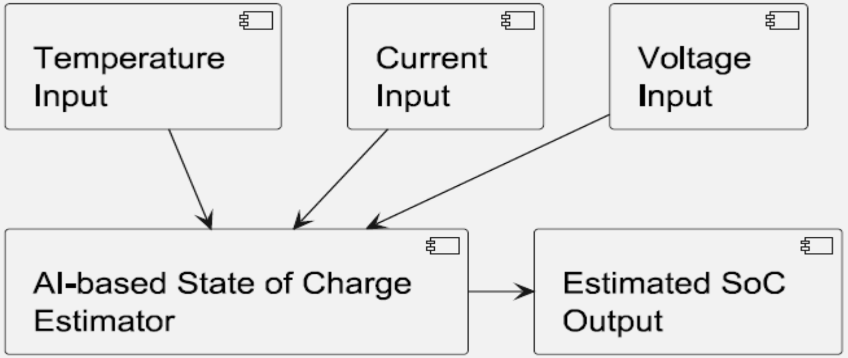
Traditional method

- Typically, a combination of methods for better accuracy
- Challenges: non-linear behavior, aging and parameter drift, individual cell differences, varying operating conditions



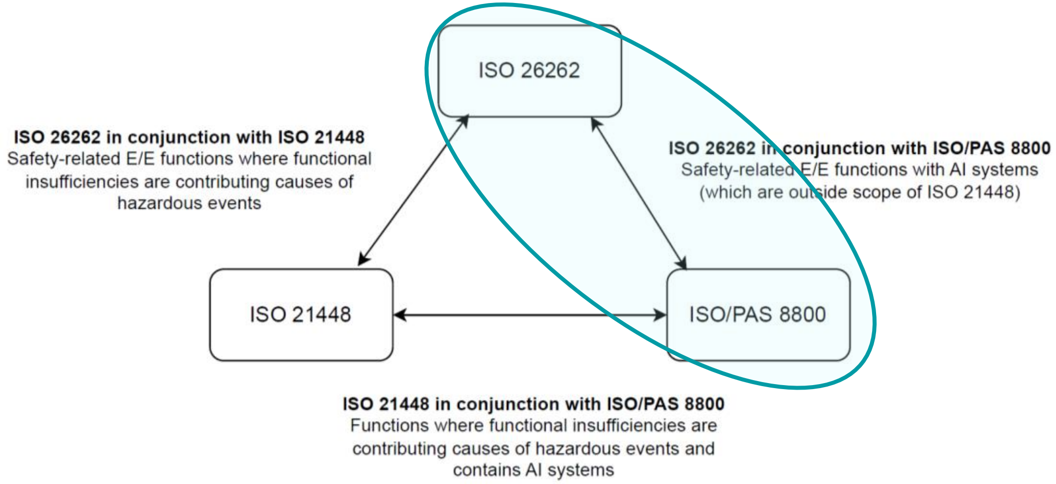
AI-based method

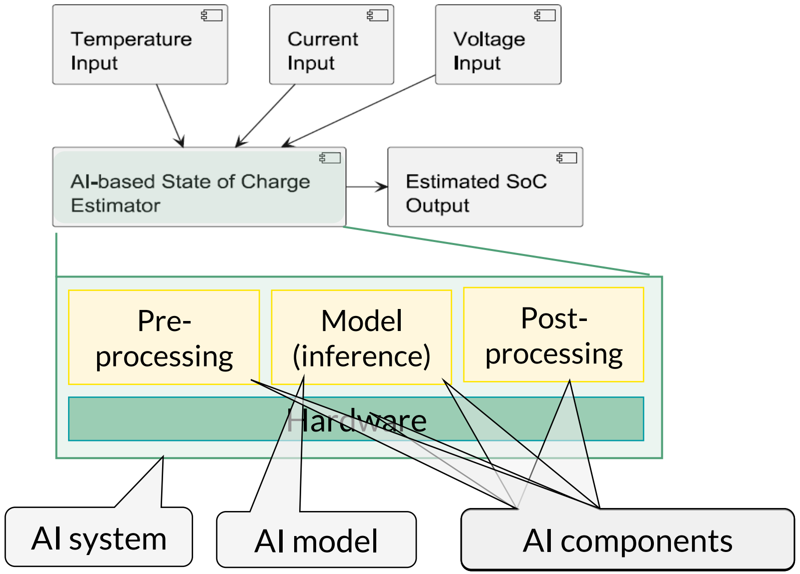
- Ability to capture the complex and non-linear behaviour, adapts to variations
- Lack of interpretability, difficult to trust for safety-critical systems



Relevant Standards for SOC Estimator

- Three main automotive safety standards
  - ISO 26262 Functional safety
  - ISO 21448 Safety of the intended functionality
  - ISO/PAS 8800 Safety and artificial intelligence
- For our SOC, use of ISO 26262 and ISO/PAS 8800

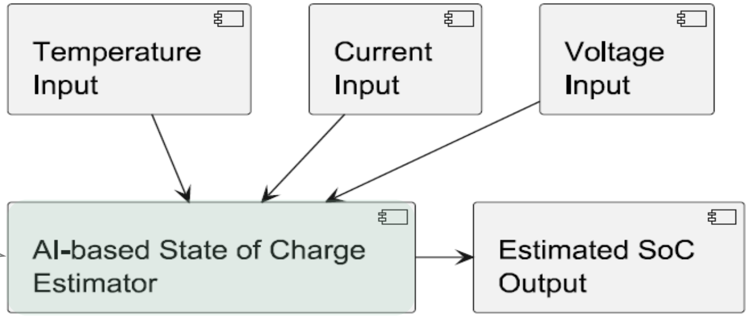




- AI components which are not an AI model developed with ISO 26262
- AI model, use of ISO/PAS 8800

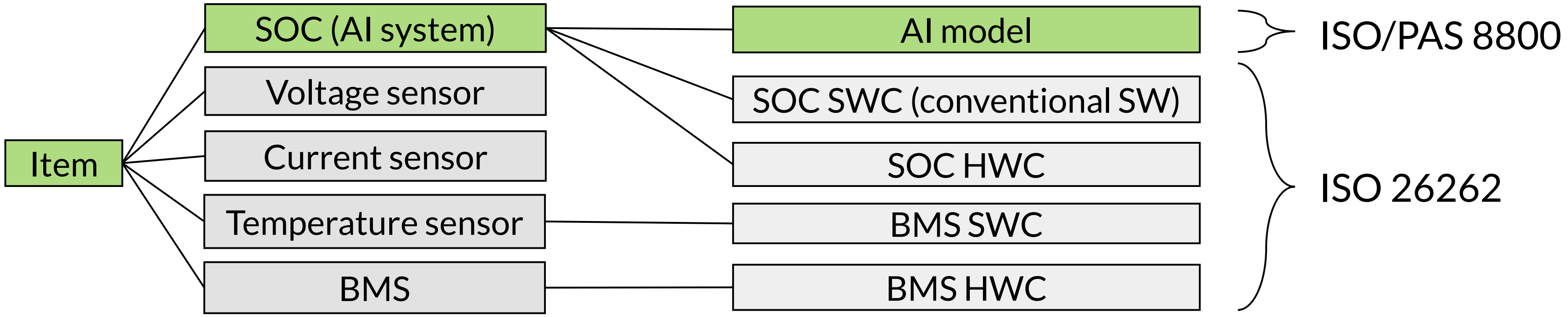
SOC Implementation

- AI-based SOC estimator from literature
  - Recurrent NN with Long Short-Term Memory that generates SOC estimations based on N preceding steps[1]
  - Parameter values with good performance for uncorrupted input were chosen
- Model was trained on an open dataset (LG 18650HG2 Li-ion Battery)[2]
- No additional safety mechanisms

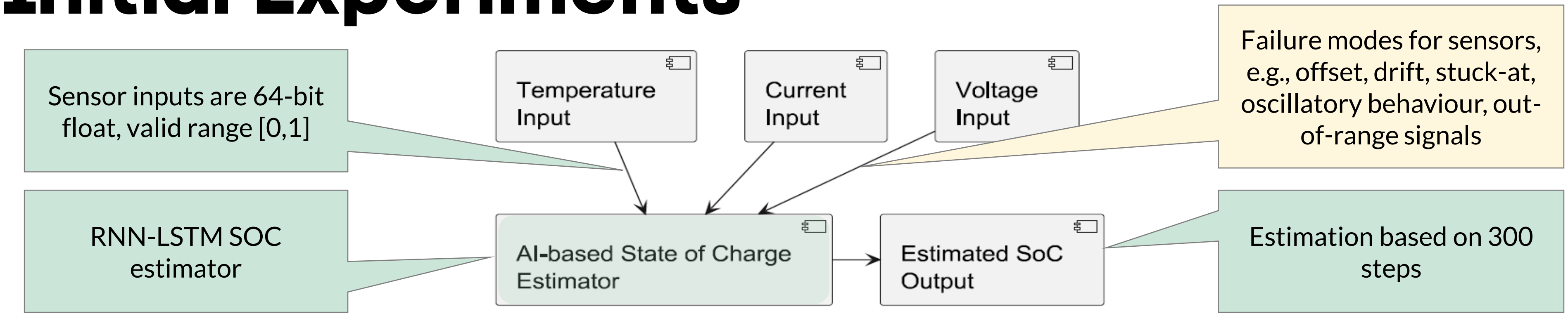


[1] K. L. Wong, M. Bosello, R. Tse, C. Falcomer, C. Rossi, and G. Pau, "Li-Ion Batteries State-of-Charge Estimation Using Deep LSTM at Various Battery Specifications and Discharge Cycles," in Proceedings of the Conference on Information Technology for Social Good, ser. GoodIT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 85–90. [Online] https://doi.org/10.1145/3462203.3475878

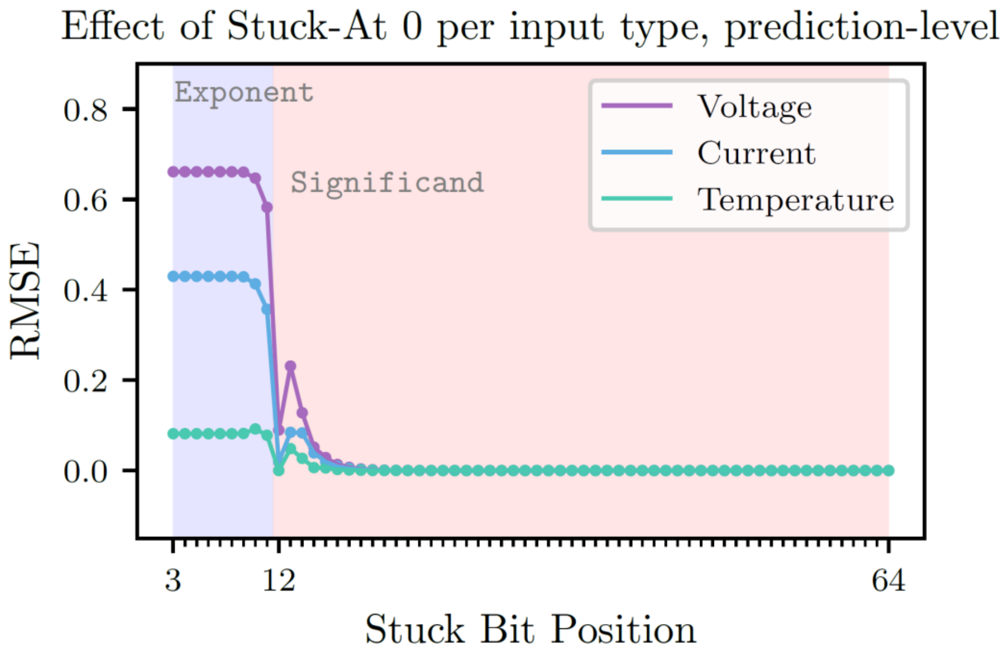
[2] P. Kollmeyer, C. Vidal, M. Naguib, and M. Skells. (2020) LG 18650HG2 Li-ion Battery Data and Example Deep Neural Network xEV SOC Estimator Script. Version 3. [Online] https://data.mendeley.com/datasets/cp3473x7xv/3

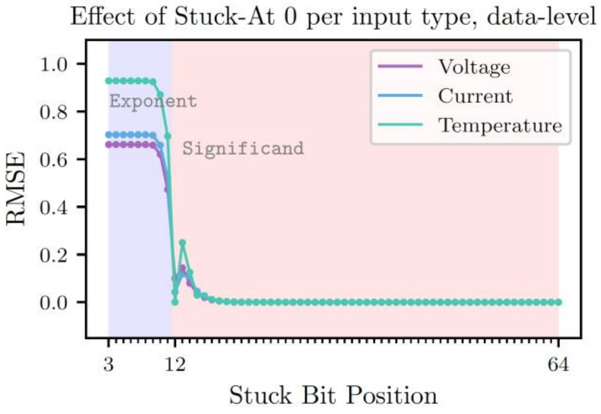


Initial Experiments



- Purpose of experiment:
  - Investigate robustness against common input (sensor) faults
  - Characterize behaviour to determine need for safety mechanisms
- First experiment: Fault-injection with stuck-at fault model for sensor inputs



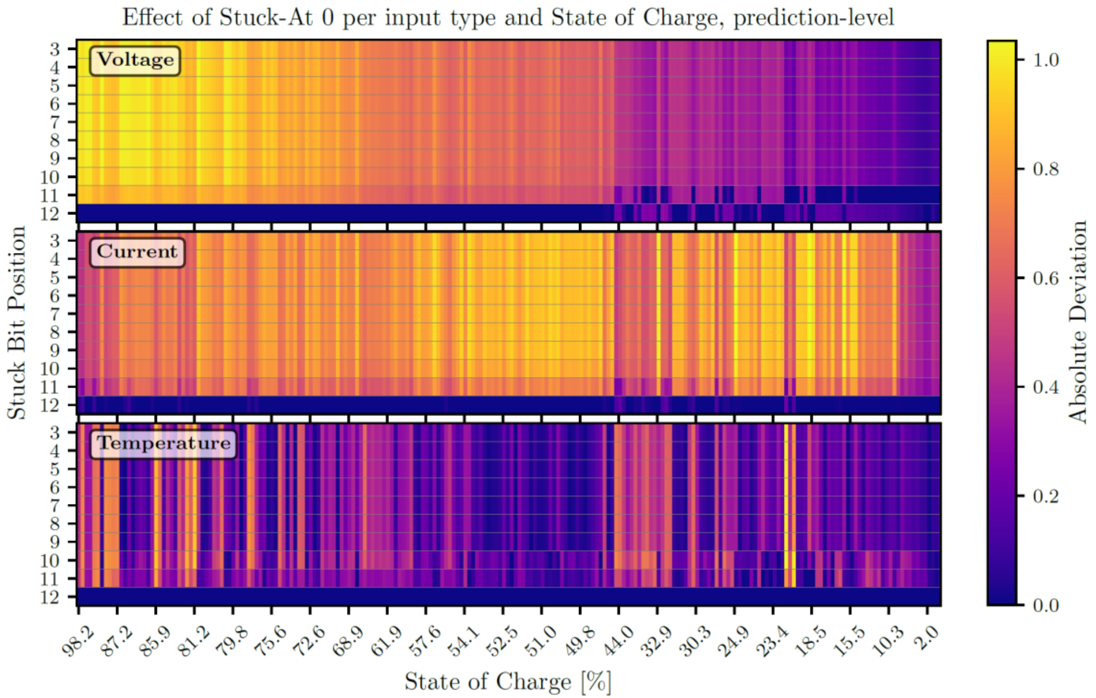


Effect of stuck-at 0

- Error (as one might expect) higher for high-value bits
- Significant difference in sensitivity between input parameters
- Error on output (prediction-level) not necessarily reflecting the most significant errors on input (data-level) side

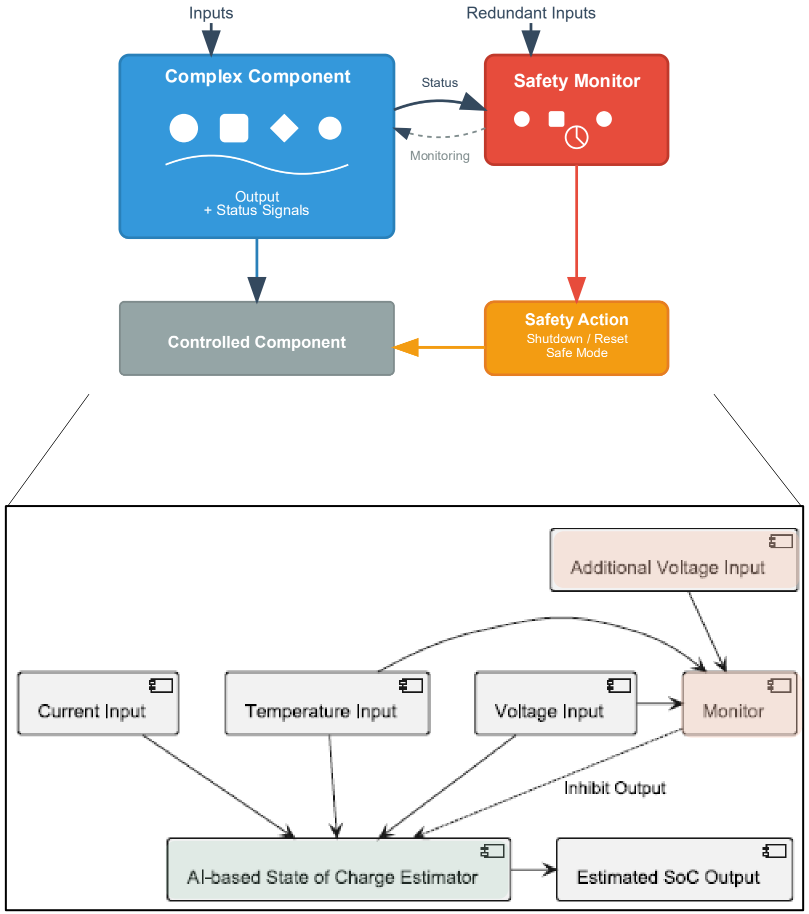
Deviation heatmap (exponent bits)

- High prediction deviation for voltage stuck-at 0 faults at high SoC → risk of overcharging



Potential Safety Mechanisms

- Safety envelope can be used for SOC
  - Guard against overcharging fault mode
  - Independence from AI SOC, conservative response
- Input range checking and/or redundant inputs
- Data augmentation
  - Expand training set to include typical sensor faults
- Adversarial training
  - Robustness against deliberate attacks
- Ensemble methods
  - Combining predictions from diverse models
- Out-of-distribution detection



Summary

- Rapidly evolving legislative and standards landscape affecting AI in critical systems
- Several existing safety assurance frameworks
  - But more experience needed
  - Example: AI-based State-of-charge estimator
- Monitoring and continuous assurance necessary for AI in safety-critical systems

RI.
SE

Dr. Fredrik Warg

Senior Researcher

Safety and Transport
Department Electrification and Dependability
Unit Dependable Transport Systems

fredrik.warg@ri.se

Research interests:
Safety assurance and V&V methods | *Connected automated vehicles* |
Safe AI | *Software engineering for dependable systems* |
Security-informed safety

@ri.se: https://www.ri.se/en/person/fredrik-warg
@orcid: https://orcid.org/0000-0003-4069-6252