# Standardisation of safe, data-driven AI Development & Tooling

KI Data Tooling – Final Event
05/06 December 2023

iso.org

# Agenda

## 01 Complexity and Uncertainty

Challenges in the use of AI for safety-critical driver assistance and automated driving tasks

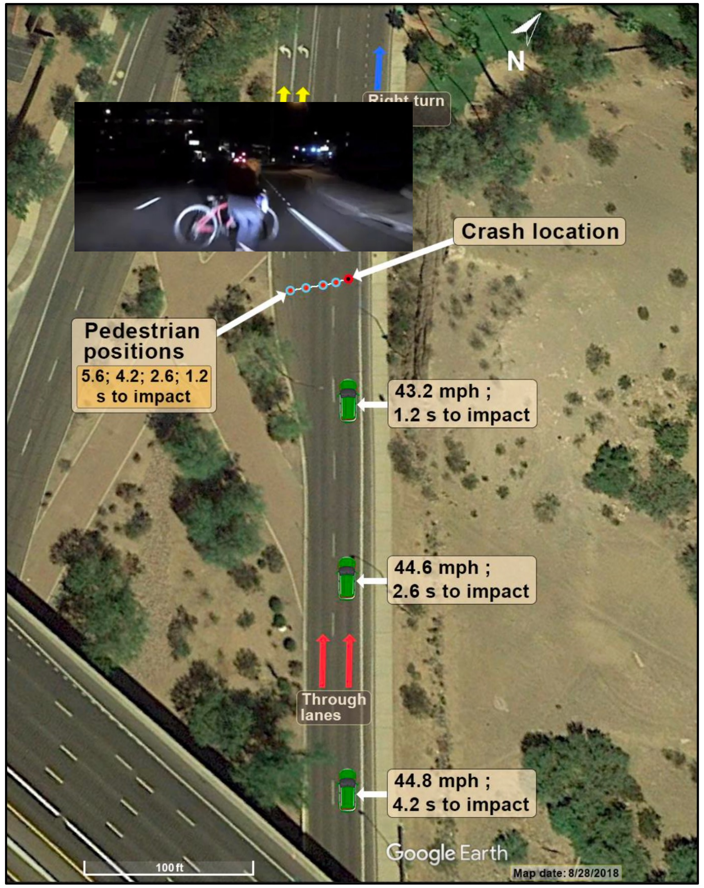
## 02 Regulations and standards for safe AI

Relevance of existing standards and recent developments (ISO PAS 8800)

## 03 What's next?

Open research questions

# Safety of complex, automated driving systems




| Time to Impact (seconds) | Speed (mph) | Classification and Path Prediction[a] | Vehicle and System Actions[b] |
|---|---|---|---|
| -9.9 | 35.1 | -- | Vehicle begins to accelerate from 35 mph in response to increased speed limit. |
| -5.8 | 44.1 | -- | Vehicle reaches 44 mph. |
| -5.6 | 44.3 | Classification: *Vehicle*—by radar<br>Path prediction: *None*; not on path of SUV | Radar makes first detection of pedestrian (classified as vehicle) and estimates speed. |
| -5.2 | 44.6 | Classification: *Other*—by lidar<br>Path prediction: *Static*; not on path of SUV | Lidar detects unknown object. Object is considered new, tracking history is unavailable, and velocity cannot be determined. ADS predicts object's path as static. |
| -4.2 | 44.8 | Classification: *Vehicle*—by lidar<br>Path prediction: *Static*; not on path of SUV | Lidar classifies detected object as *vehicle*; this is a changed classification of object and without a tracking history. ADS predicts object's path as static. |
| -3.9[c] | 44.8 | Classification: *Vehicle*—by lidar<br>Path prediction: Left through lane (next to SUV); not on path of SUV | Lidar retains classification *vehicle*. Based on tracking history and assigned goal, ADS predicts object's path as traveling in left through lane. |
| -3.8 to -2.7 | 44.7 | Classification: alternates between *vehicle* and *other*—by lidar<br>Path prediction: alternates between *static* and left through lane; neither considered on path of SUV | Object's classification alternates several times between *vehicle* and *other*. At each change, tracking history is unavailable; ADS predicts object's path as static. When detected object's classification remains same, ADS predicts path as traveling in left through lane. |
| -2.6 | 44.6 | Classification: *Bicycle*—by lidar<br>Path prediction: *Static*; not on path of SUV | Lidar classifies detected object as *bicycle*; this is a changed classification of object and object is without a tracking history. ADS predicts bicycle's path as static. |
| -2.5 | 44.6 | Classification: *Bicycle*—by lidar<br>Path prediction: Left through lane (next to SUV); not on path of SUV | Lidar retains *bicycle* classification; based on tracking history and assigned goal, ADS predicts bicycle's path as traveling in left through lane. |

Source: National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian Tempe, Arizona march 18, 2018. 2019.

**Governance** — Failure to regulate accountability for safety of automated driving

**Management** — Inadequate engineering and operating processes, lack of oversight of safety driver

**Human factors** — Failure of safety driver to detect that system was not operating correctly

**Technical** — Failure of system to correctly detect pedestrian and avoid collision

Burton, Simon, John Alexander McDermid, Philip Garnett, and Rob Weaver. "Safety, Complexity, and Automated Driving: Holistic Perspectives on Safety Assurance." Computer 54, no. 8 (2021): 22-32.

We need to acknowledge the inherent complexity of the task, environment and system...

... and its impact on our ability to provide convincing safety assurance arguments

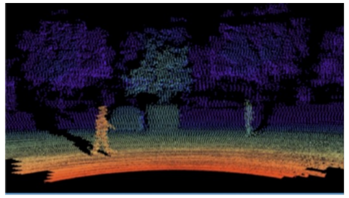
# The impact of complexity

## Complexity inevitably leads to ***uncertainty in the safety assurance argument***

***Uncertainty:*** *Any **deviation** from the unachievable ideal of **completely deterministic knowledge** of the relevant system**

Source: https://www.bbc.com/news/world-asia-india-38155635

**Scope & unpredictability** of operational domain and critical events



Source: https://velodynelidar.com

**Inaccuracies & noise** in environmental sensors and signal processing

Source https://www.cityscapes-dataset.com/examples

**Heuristics or machine learning techniques** with unpredictable results

Uncertainty in our understanding of the **environment** and **task**

Uncertainty in whether our **observations** of the environment are accurate and complete

Uncertainty in how our **system** (especially AI/ML) processes inputs and makes decisions

*W. E. Walker et al. "Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support". In: Integrated Assessment 4.1 (Mar. 2003), pp. 5–17. ISSN: 1389-5176. DOI: 10.1076/iaij.4.1.5.16466.

# Existing Standards – Functional Safety (ISO 26262)

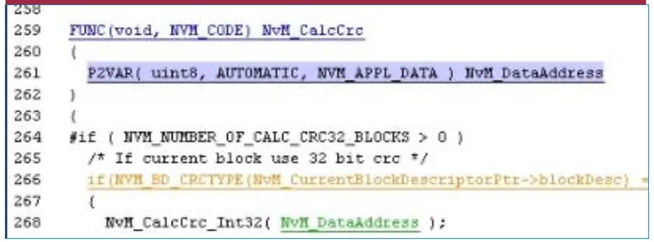
Absence of unreasonable risk due to hazards caused by **malfunctioning behaviour** of E/E systems

Risk associated with malfunctioning behaviour

Random hardware errors


Photo: Christian Taube - Own work

Systematic errors (HW and SW)

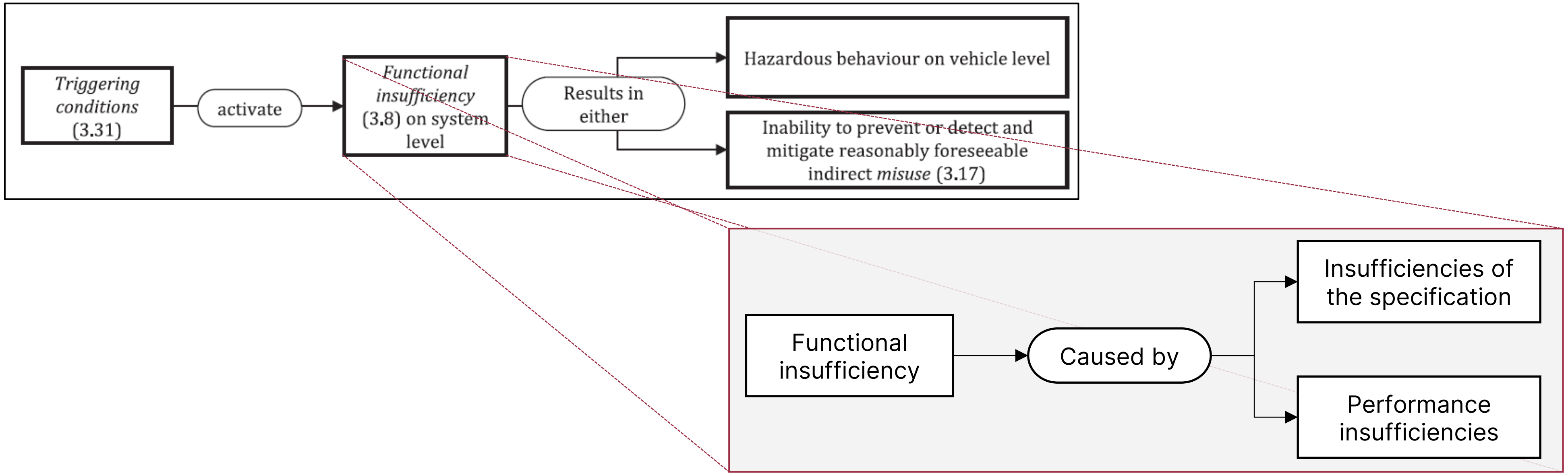



Picture: Mathworks

...is a pre-requisite for AI/ML-based automated driving systems

# Existing Standards – Safety of the intended functionality (ISO 21448)

Absence of unreasonable risk due to hazards resulting from **functional insufficiencies** of the intended functionality or by reasonably foreseeable misuse by road users




...interpretation and operationalization required for AI/ML-based systems

# Insufficiencies of the specification
# An ML interpretation

- Developing a complete set of safety requirements
  - How to demonstrate the completeness of requirements for an inherently complex task?
  - Which KPIs/Metrics can be used to measure the conformance to the requirements?
  - How to derive target values (validation targets) for these metrics?
- **Data as the specification**
  - How to argue the integrity and appropriateness of the data?
  - How to demonstrate coverage of the operational domain and requirements?
- Requires a detailed understanding of the operational domain and technical system context
  - How to deal with rare but critical events ?
  - How to deal with distributional shift / changes in the environment over time?

# Performance insufficiencies
# An ML interpretation

- Machine learning: Optimizes model parameters through computational techniques, such that the model's behaviour reflects the training data (as an approximation of the target function)
- Performance insufficiencies of ML: gaps between theoretically optimal function and the trained model:
  - Characterized as lack of generalization and robustness, bias, etc.
  - Related to the concepts of task complexity/learnability, sample complexity and model expressiveness
- How to ensure the model meets its requirements and demonstrate this with a sufficient level of confidence?
  - **Which verification data to use?**
  - Exacerbated by further properties of ML models such as lack of explainability and prediction uncertainty

# Emerging standards for Safe AI

ISO/IEC JTC 1/SC 42
Information Technology – Artificial intelligence

ISO/TC 22/SC 32 Road Vehicles – Electrical and electronic components and general system aspects

Narrow – Scope - Broad

## AI

ISO/IEC 22989 – AI Definitions

ISO/IEC TR 24028 – AI Trustworthiness

ISO/IEC TR 24029 – NN Robustness,
ISO/IEC DTS 4213.2* ML Classification performance,
....

## Safety

IEC 61508 – Functional safety of E/E systems

ISO/IEC TR 5469* – Functional safety & AI Systems

## Road vehicle specific

ISO 26262 – Functional safety

ISO 21448 – Safety of the intended functionality

**ISO PAS 8800* – Safety & AI**

ISO TR 4804 / TS 5083* – Safety for Automated Driving Systems

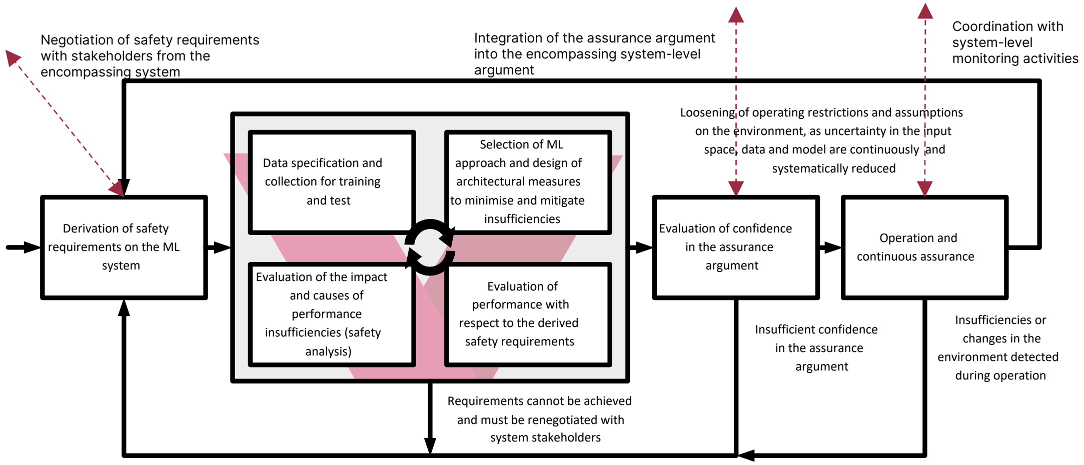
* Under development

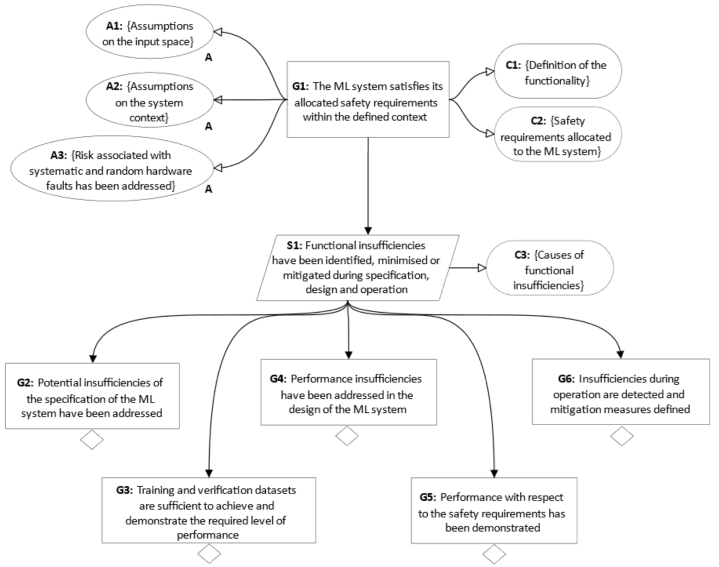
**ISO PAS 8800:**

- Operationalization of SOTIF concepts for AI/ML-based vehicle functionality, ISO 26262 as pre-requisite
- Not restricted to automated driving functions or specific ML techniques

# Concepts of ISO PAS 8800

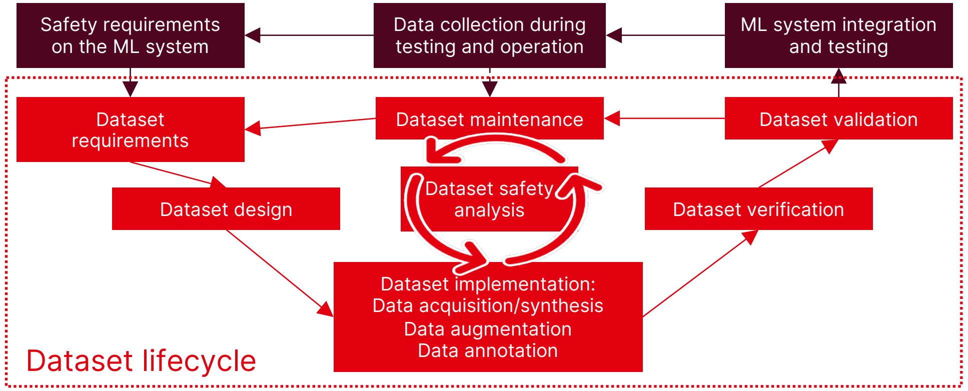
- Definition of a **fault model** and **safety-related properties** used to define detailed safety requirements of the AI systems
- Definition of an **iterative AI safety lifecycle**, including continuous safety assurance during operation
- Definition of **development** and **architectural** measures for achieving the safety-related properties of AI systems
- Definition of a **Data-lifecycle** and associated **safety-related data properties**
- **Safety analyses** and **structured assurance arguments** for justifying acceptable residual risk associated with the AI system



# ISO PAS 8800 – Data related considerations

- **Dataset lifecycle:**
  - Apply a systematic approach to the gathering, creation, analysis, verification, validation management and maintenance of datasets used in the development of the ML system
  - Identify which properties of the datasets have an impact on the safety requirements of the ML system
- **Dataset safety analysis:**
  - Identify dataset errors that may impact the safety requirements
  - Define measures to prevent or mitigate these errors
- Requires **application-specific interpretation** and consideration of the **integrity** and **sufficiency of the data**



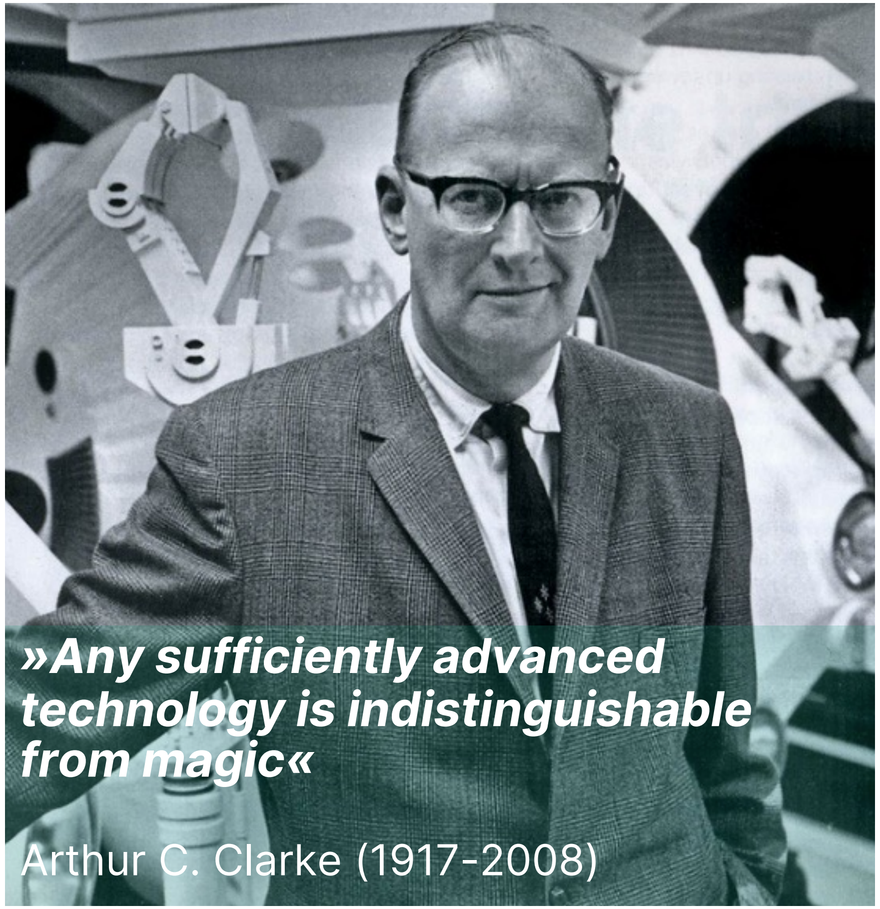

| Common dataset errors |
| --- |
| Lack of coverage of the input space |
| Lack of representation of safety-relevant edge cases |
| Distribution does not match the target input space |
| Dependencies on the data acquisition method (e.g. camera type, geographic, temporal dependencies) |
| Data fidelity (e.g., sensor noise, accuracy of synthetic data) |
| Errors in the meta-data / labelling |
| Lack of independence between training and verification datasets |

# What’s next

- Safety assurance of ML-based safety-relevant functions requires managing complexity and uncertainty in:
  - The task and environment
  - Data
  - and the system (from sensors to ML-models)
- First generation of standards and regulations will provide guidance on important principles for achieving an acceptable level of residual risk…
- …but will require (a lot of) application and ML technology-specific interpretation
- For many realistically complex tasks, an appropriate combination of safety assurance methods have yet to be found

# Open research questions

- **How safe is safe enough?**
  - Defining the Operational Design Domain as a basis for design and test
  - Operationalizing abstract requirements into measurable properties
- **Engineering safe AI/ML-based systems**
  - **Safety-grade datasets with demonstrable properties**
  - Selection and optimization of AI/ML approaches for safety-critical perception and planning tasks
  - Analysing the impact of uncertainty within the system
  - Design of monitoring and redundancy measures for compensating for uncertainty in sensors and AI components
- **Arguing the safety of AI/ML-based systems**
  - V&V of perception and planning functions
  - Continuous, automated safety assurance
  - Demonstrating confidence in evidence and assurance arguments



# Thank you.

Making lives ***easier***, ***safer*** and ***better***.

**Prof. Dr. Simon Burton**
Convenor ISO TC22/SC32/WG14 – Safety and AI
Honorary Visiting Professor, University of York
simon.burton@safer-complex-systems.de