Centre *for* Assuring Autonomy

UNIVERSITY *of York*

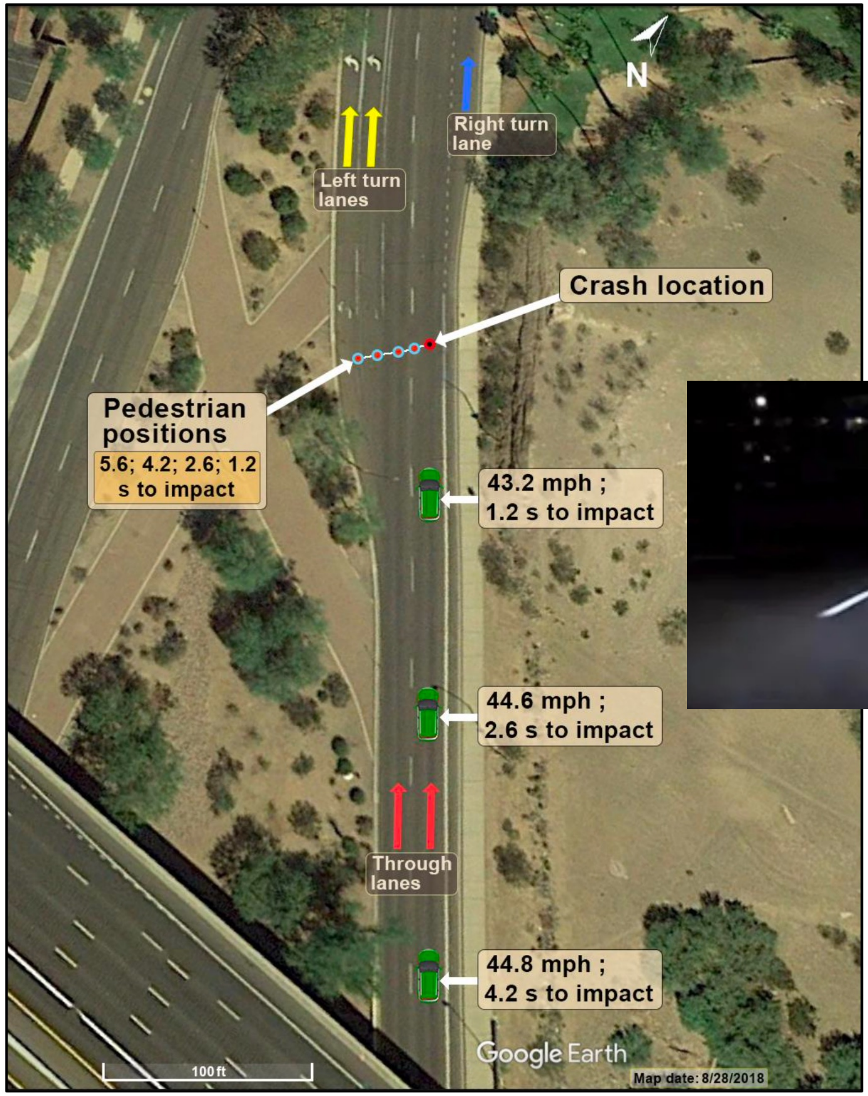
Automotive standards for AI safety and research perspectives

Safety under uncertainty

Prof. Simon Burton,
Chair of Systems Safety, Business Director - CfAA

LR Foundation

Automotive safety and AI



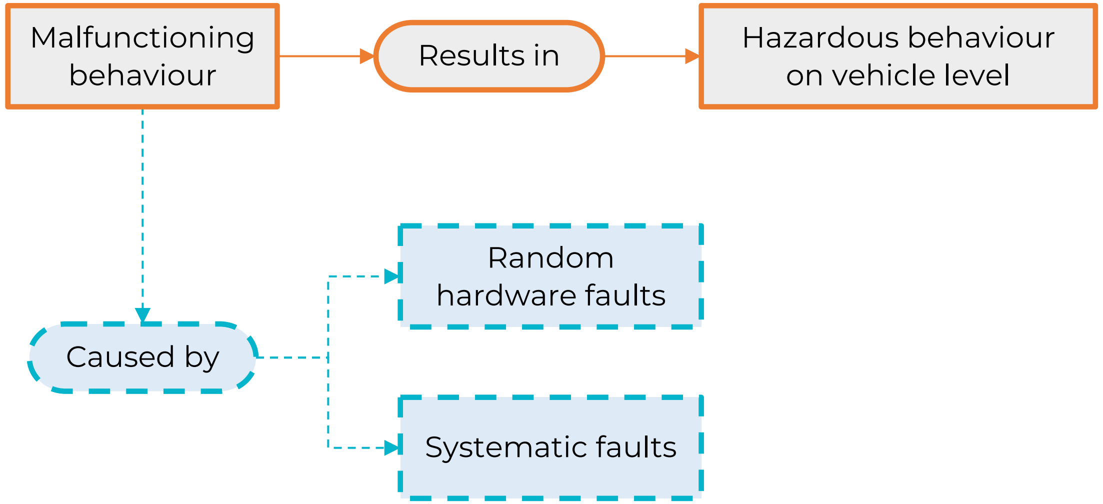
Time to Impact (seconds)	Speed (mph)	Classification and Path Prediction[a]	Vehicle and System Actions[b]
-9.9	35.1	--	Vehicle begins to accelerate from 35 mph in response to increased speed limit.
-5.8	44.1	--	Vehicle reaches 44 mph.
-5.6	44.3	Classification: *Vehicle*—by radar Path prediction: *None*; 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 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.
[illegible]	4.8	Classification: *Vehicle*—by lidar 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.
[illegible]	4.8	Classification: *Vehicle*—by lidar 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.
[illegible]	4.7	Classification: alternates between *vehicle* and *other*—by lidar 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 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 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.

Wider context of automotive safety standards

ISO 26262: Functional safety

“Absence of unreasonable risk due to hazards caused by ***malfunctioning behaviour*** *of the electrical and/or electronic systems”*



Also addresses:

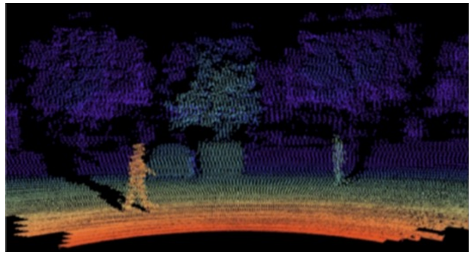
- Safety management (organisational and project-specific)
- Supporting processes

Safety challenges of automated driving functions

Impact of environment, task and system complexity

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

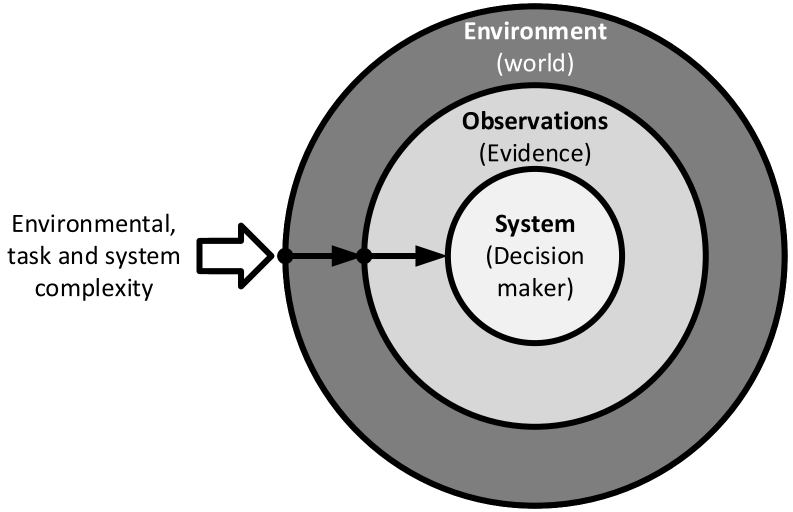
Scope & unpredictability of operational domain and critical events



Inaccuracies & noise in environmental sensors and signal processing

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

Heuristics or machine learning techniques with unpredictable results



Manifestations of uncertainty

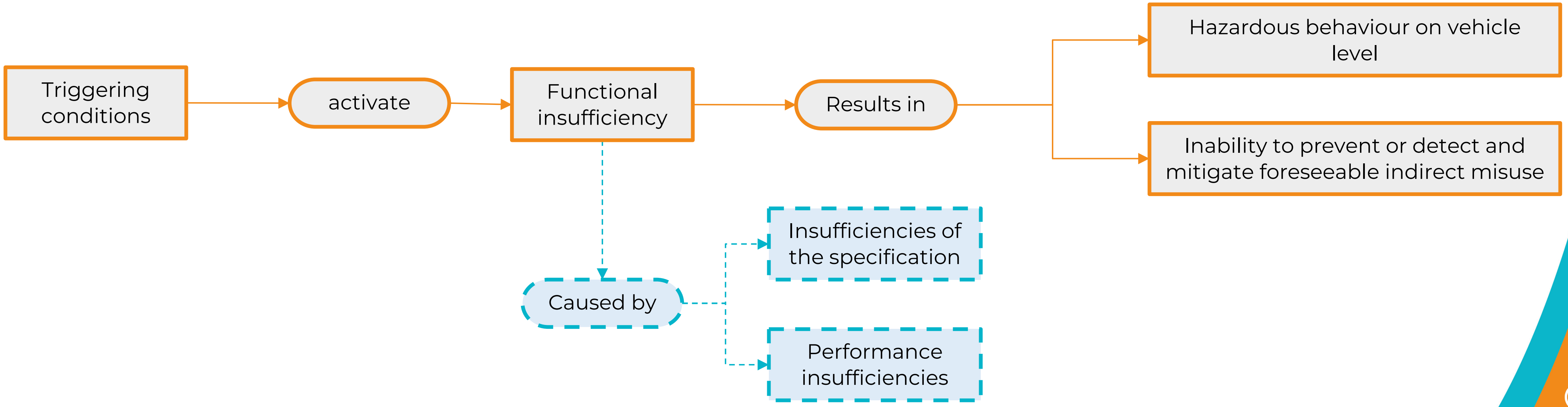
Burton, Simon, and Benjamin Herd. "Addressing uncertainty in the safety assurance of machine-learning." *Frontiers in Computer Science* 5 (2023), Inspired by: Lovell, B. E. (1995). A Taxonomy of Types of Uncertainty. Portland State University.

Centre *for* Assuring Autonomy

Wider context of automotive safety standards

ISO 21448: Safety of the intended functionality (SOTIF)

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



Centre *for* Assuring Autonomy

Safety challenges of AI-based functions

Insufficiences of the specification

How to define a "complete" specification:

- Dealing with **rare but critical events**
- **Distributional shift** / changes in the environment over time
- Requires a detailed understanding of the operational domain and technical system context
- 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 demonstrate coverage of the operational domain and requirements?
- Does the (ground truth) data accurately represent the intended functionality for all possible scenarios?

Safety challenges of AI-based functions

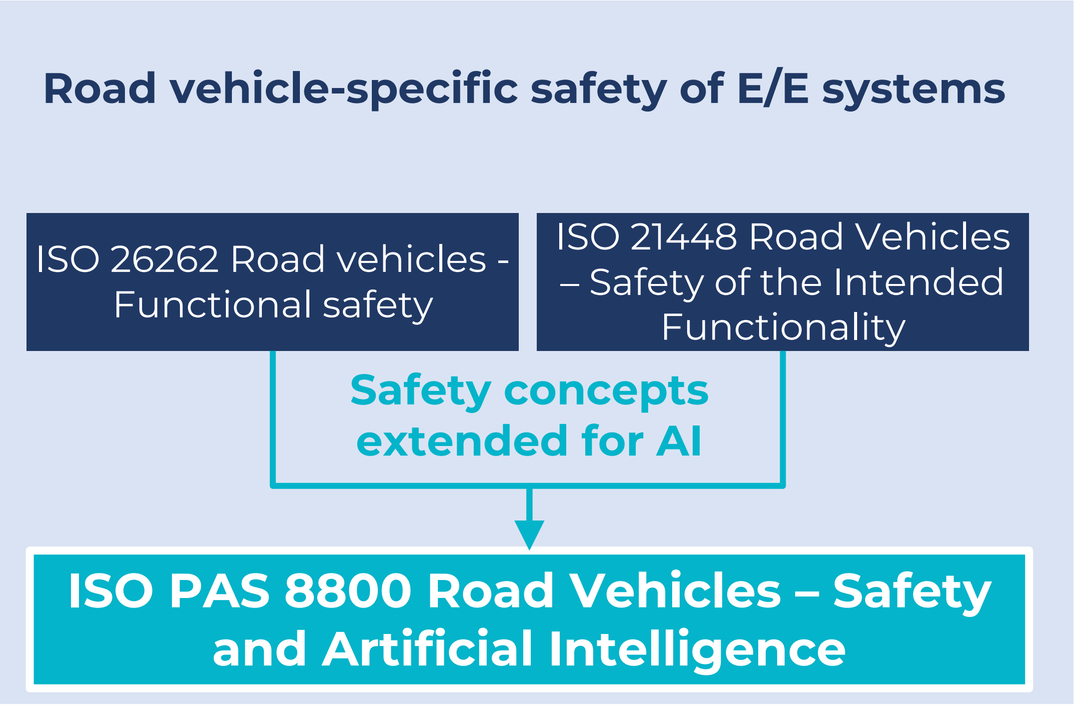
Performance insufficiencies

Model uncertainty:

- **Residual errors:** due to bias and lack of generalization and robustness: outputs sensitive to small changes in the inputs and insufficiencies in training data
- **Prediction uncertainty**: Confidence scores not necessarily indication of probability of correctness
- Related to the concepts of task complexity, sample complexity and model expressiveness
- How to systematically identify triggering conditions and demonstrate a lack of "unknown triggering" conditions?

Centre *for* Assuring Autonomy

ISO PAS 8800


Road vehicle-specific safety of E/E systems
ISO 26262 Road vehicles - Functional safety
ISO 21448 Road Vehicles – Safety of the Intended Functionality
Safety concepts extended for AI
ISO PAS 8800 Road Vehicles – Safety and Artificial Intelligence

Overview of ISO PAS 8800

Scope

- Extension of concepts from ISO 26262 and ISO 21448
- Process oriented standard based on a safety-lifecycle
- Only a few high-level requirements defined for each lifecycle phase
  - Not specific to a particular AI/ML technology
  - However, most recommendations and examples oriented towards machine learning
  - Not specific to particular applications (e.g. automated driving)
- Informative guidance to serve as an interpretation aid of the requirements and not necessarily to promote specific solutions

Through-life assurance

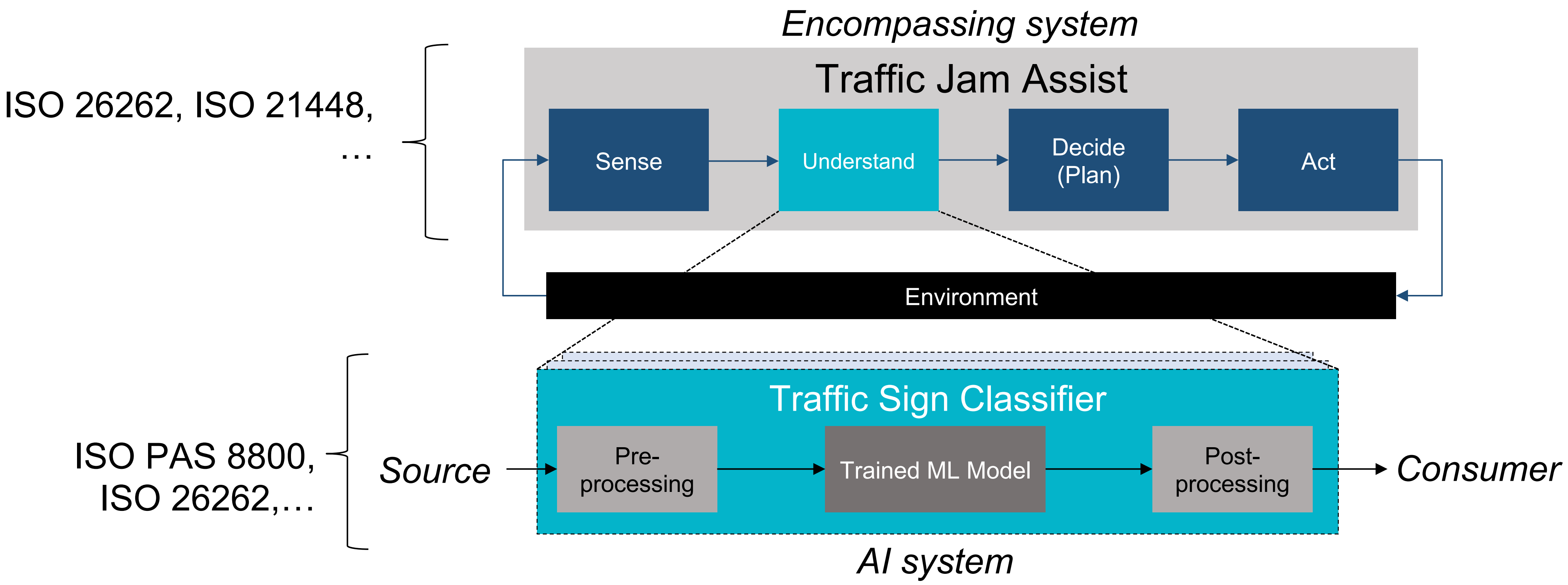
AI system:
Pre- and postprocessing to reduce impact of AI errors, consideration of known insufficiencies in system requirements, assurance argument

AI model:
Specification of safety related (quantitative) properties, measures to reduce technical uncertainty, V&V, Safety Analysis

Centre *for* Assuring Autonomy

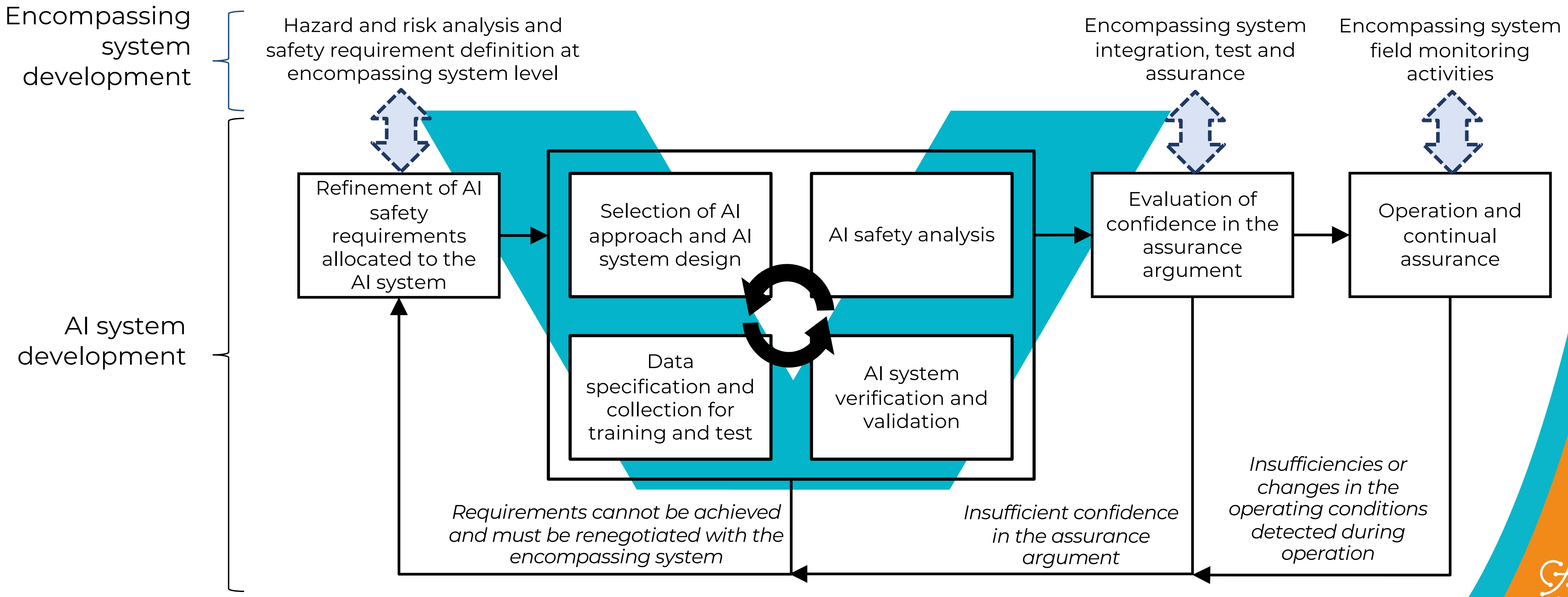
Overview of ISO PAS 8800

Example scoping of the standards



Overview of ISO PAS 8800

AI Safety lifecycle



Centre for Assuring Autonomy

Overview of ISO PAS 8800

Derivation of safety requirements (Example)

Safety requirement

Correctly classify construction signs for any given image

Property	Derived requirements
Generalization	The TSC shall achieve a high recall rate for construction signs
Robustness	The TSC should be robust against camera noise
	The TSC should be robust against partial occlusion of or damage to the traffic sign
Bias	For each combination of possible weather and lighting conditions
Prediction uncertainty	The confidence scores shall be representative of the probability of failure
…	…

Acceptance criteria

$< 10^{-04}$ missed detections/construction sign

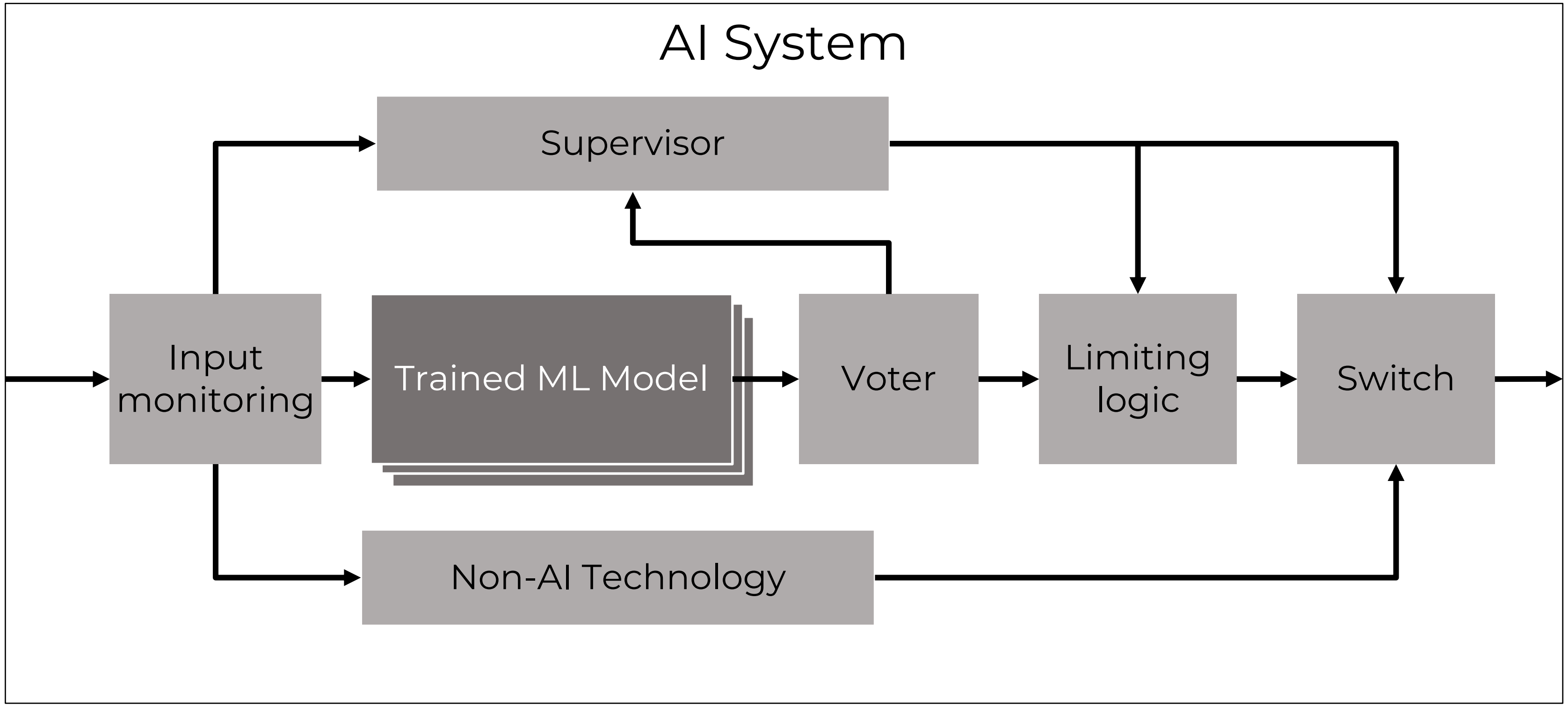
Metrics / Targets
Recall 99.99%
Adding noise perturbations characterized by $L1$norm < 0.001 on the image, shall introduce at most 0.01% false negatives
Occlusion of the traffic sign of 25% shall introduce at most 0.01% false negatives
Recall of 99.99% shall be achieved for all equivalence classes of weather and lighting
Maximum Calibration Error < 0.01
…

In addition, the limitations of the AI model and AI system must be characterized so that these can be compensated for at the level of the encompassing system

Centre *for* Assuring Autonomy

Overview of ISO PAS 8800

Design concepts

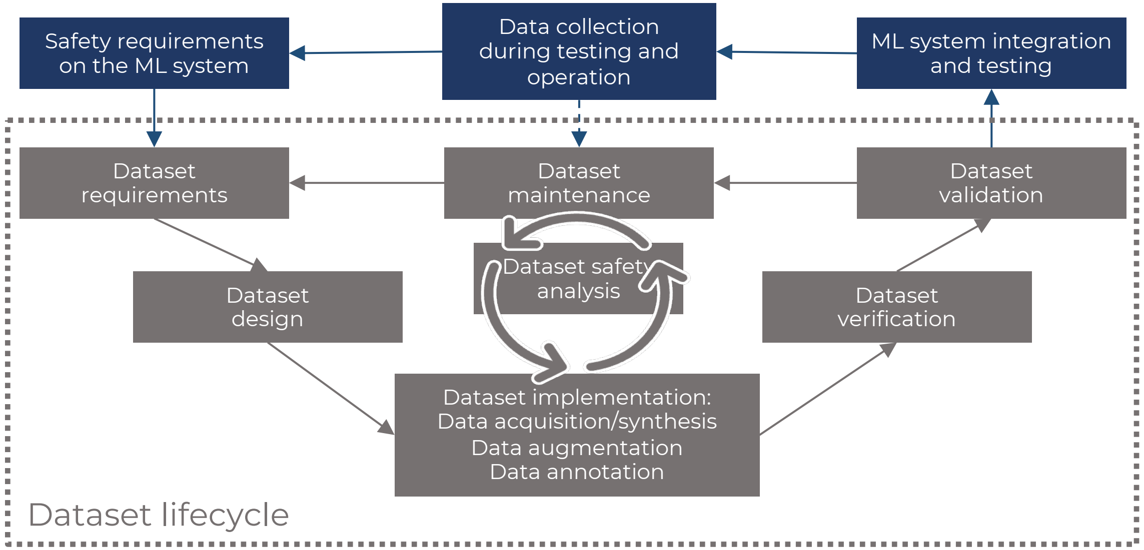


Can help to reduce the absolute performance requirements on the ML model by compensating for residual errors

Centre *for* Assuring Autonomy

Overview of ISO PAS 8800

Data lifecycle and dataset safety analysis



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

Centre *for* Assuring Autonomy

Overview of ISO PAS 8800

Verification, Validation and Safety Analysis:

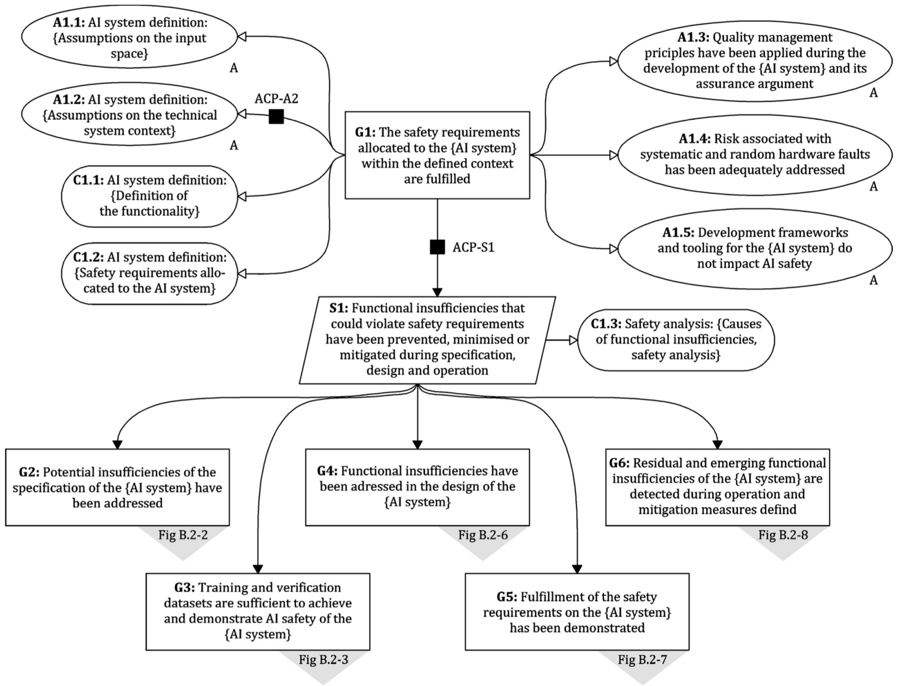
- Limited transferability of software verification techniques
- Increased reliance on statistical and search-based testing
- Virtual testing vs. physical testing
- Safety analysis
    - A direct relationship between causes of errors and their consequences may be difficult to determine/disentangle.
    - An evaluation of the effectiveness of proposed measures is therefore essential.

Centre *for* Assuring Autonomy

Overview of ISO PAS 8800

Safety assurance argument

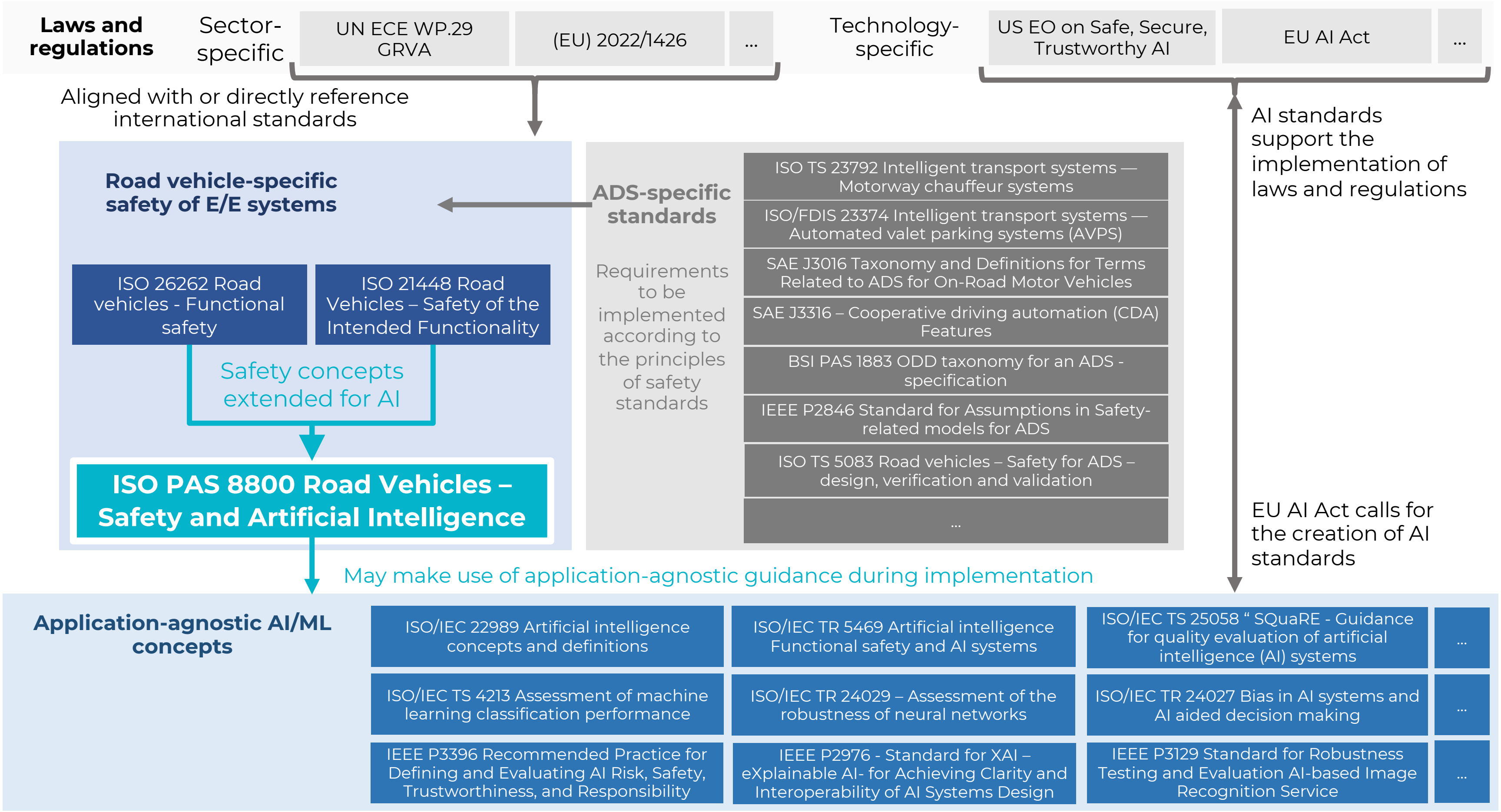
- Develop an assurance argument demonstrating that the AI safety requirements are fulfilled
- As a contribution to the safety assurance argument of the encompassing system
- Continually re-evaluated and updated during operation



Centre for Assuring Autonomy

Wider context of automotive safety standards

A complex evolving landscape of standards and regulation



Centre for Assuring Autonomy

Safety under uncertainty

Centre *for* Assuring Autonomy

Safety under uncertainty

Principles of effective assurance arguments*

- Clear definition of the safety claim to be demonstrated
  🤔 How to formulate safety requirements as measurable properties of ML models?
- Assurance driven workflow for continually/incrementally capturing evidence during development and operation
  👍🏻 Covered by ISO PAS 8800 and other standards
- Arguments based on rigorous models of the system and its context
  🤔 Opaque models/ML explainability, incomplete definition of the input space?
- Use of evidence and arguments that can be easily refuted or believed
  🤔 Can we trust our ML metrics to provide us with an accurate evaluation of safety risk?

*With thanks to Natarajan Shankar, SRI: Keynote SAFECOMP 2023

Safety under uncertainty

Ongoing research

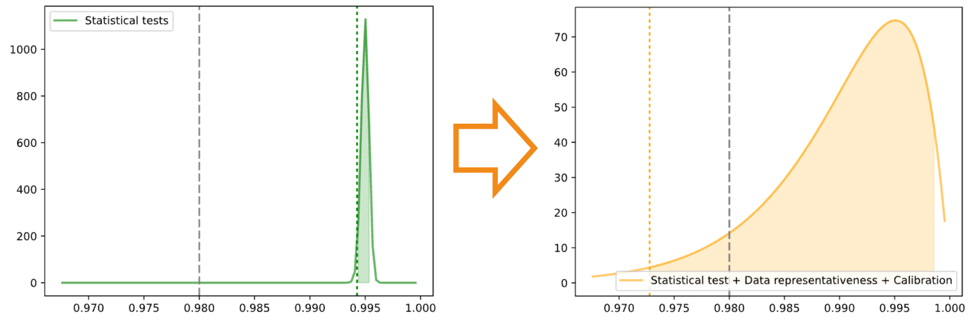
Many metrics are proposed for evaluating the safety of ML-based functions, do they really provide a realistic estimation of the actual safety risk?

Estimated safety **Actual safety**

$$\frac{\#\{j \in I : A(j) \wedge P(j, M(j))\}}{\#\{j \in I : A(j)\}} \approx \frac{\sum_{i \in I, A(i) \wedge G(i, M(i))} \mathbb{P}_{ODD}(i)}{\sum_{i \in I, A(i)} \mathbb{P}_{ODD}(i)}$$

Assurance Uncertainty

1. Collect primary evidence to directly support the safety claim including uncertainty
2. Identify evidence to support or refute the validity of the primary evidence
3. Adjust estimates of safety risk based on uncertainty in the measurement

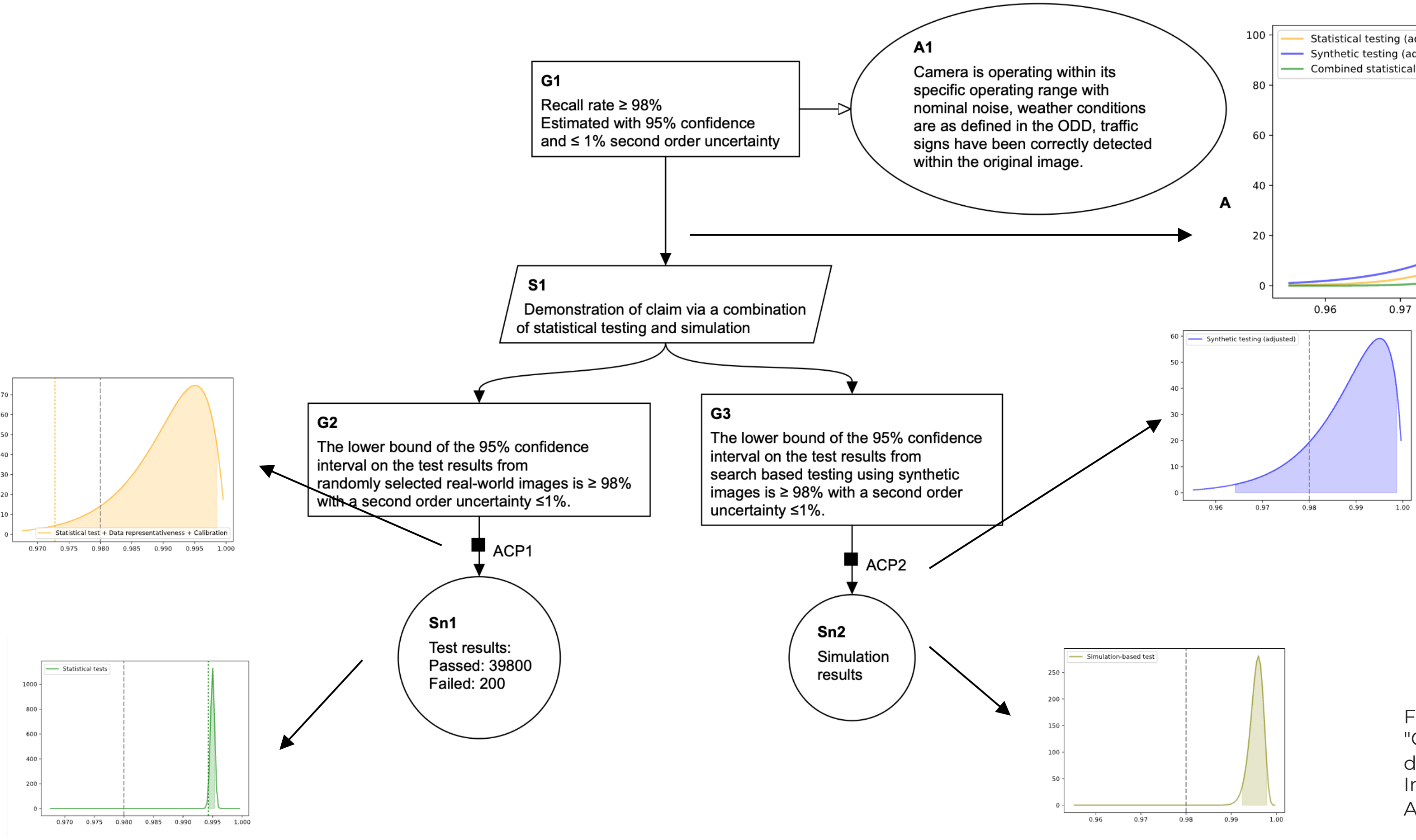


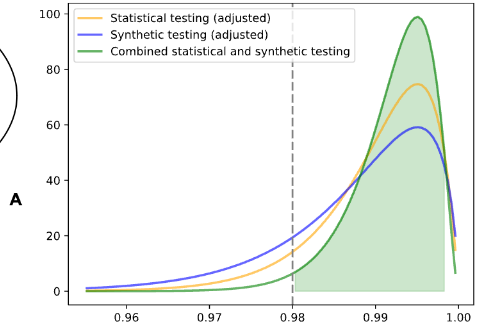
For more details see: Herd, Benjamin, and Simon Burton. "Can you trust your ML metrics? Using Subjective Logic to determine the true contribution of ML metrics for safety." In Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, pp. 1579-1586. 2024.

Centre *for* Assuring Autonomy

Assurance uncertainty

Uncertainty aware safety arguments





Combined evidence ω_{com}	
Adj. Expected value	99.2%
Adj. 95% credible interval lower bound	98.0%
2nd Order Uncertainty	1.2%

For more details see: Herd, Benjamin, and Simon Burton. "Can you trust your ML metrics? Using Subjective Logic to determine the true contribution of ML metrics for safety." In Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, pp. 1579-1586. 2024.

Centre *for* Assuring Autonomy

Conclusions and next steps

Centre *for* Assuring Autonomy

Conclusions and next steps

Research: Foundations of convincing AI safety arguments

Convincing arguments for AI safety require:

- A precise definition of the properties being measured and their relationship to system requirements
    - Safety requirements → Measurable properties
- Evidence beyond simple metrics calculated based on arbitrary test data
    - Rigorous approach to statistical reasoning based on quantitative evidence
- Reducing uncertainty in the integrity and validity of evidence
    - Advancing state-of-the-art in (virtual) testing of AI-based systems
    - Scaling formal verification of well-bounded properties such as robustness
- High integrity safety measures at the architectural level to mitigate against residual errors in the model
    - Balancing safety risk against utility (overly restrictive safety measures)

Conclusions

Summary

- Initial standards define AI safety lifecycles and iterative approaches to collecting and evaluating evidence
- The ability to provide a convincing argument for the safety of AI-based autonomy is inherently linked to the complexity of the environment, the task and the resulting models.
- Acknowledgement and management of the resulting uncertainties is required to make a convincing safety argument.
- The greater the complexity of the environment, task and system (AI models), the harder it is to trust the evidence, the assumptions and the argument structure itself.
- This may lead to the need for inherently resilient (and anti-fragile) systems, which are not fully assured in a classical sense during development.

Centre *for* Assuring Autonomy

UNIVERSITY of York

Thank you for your attention, any questions?

www.york.ac.uk/assuring-autonomy
assuring-autonomy@york.ac.uk
Centre for Assuring Autonomy LinkedIn

LR Foundation