SCANDINAVIAN CONFERENCE ON SYSTEM & SOFTWARE SAFETY

SAFETY, COMPLEXITY, AI AND AUTOMATED DRIVING
HOLISTIC PERSPECTIVES ON SAFETY ASSURANCE

SIMON BURTON, FRAUNHOFER IKS

23RD NOVEMBER 2021

Autonomous systems are inherently complex

ISO 26262: FUNCTIONAL SAFETY

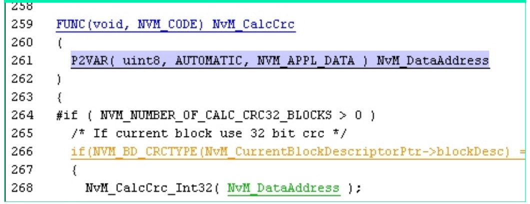
*Absence of unreasonable **risk** due to hazards caused by **<u>malfunctioning behaviour</u>** of E/E systems*

Risk associated with malfunctioning behaviour

Random Hardware Errors

Photo: Christian Taube - Own work

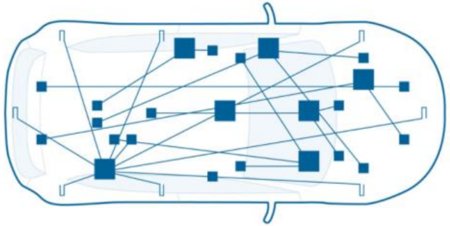
Systematic errors (HW and SW)



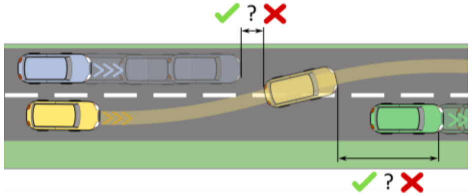
Picture: Mathworks

WHATS CHANGING? SYSTEM COMPLEXITY AND UNCERTAINTY

https://www.bosch-mobility-solutions.com/en/mobility-topics/ee-architecture/



Increasing complexity of E/E Architectures

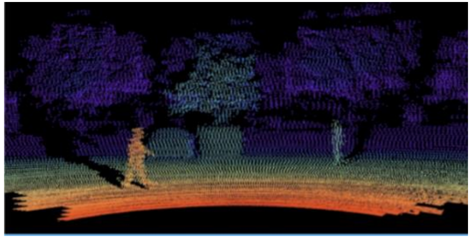


Complex behavioural interactions between systems

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

Self-organization and ad-hoc systems-of-systems

Scope & unpredictability of operational domain and critical events



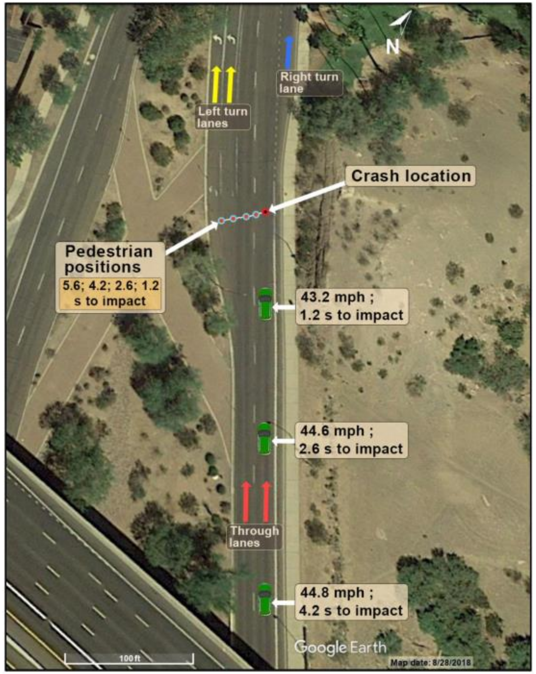
Source: https://velodynelidar.com

Inaccuracies & noise in environmental sensors and signal processing

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

Heuristics or machine learning techniques with unpredictable results

MORE THAN JUST A TECHNICAL CHALLENGE



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

Failures

Governance: Failure to regulate accountability for safety of automated driving

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

Interaction: Failure of safety driver to detect that system was not operating correctly

Technical: Failure of system to correctly detect pedestrian and avoid collision

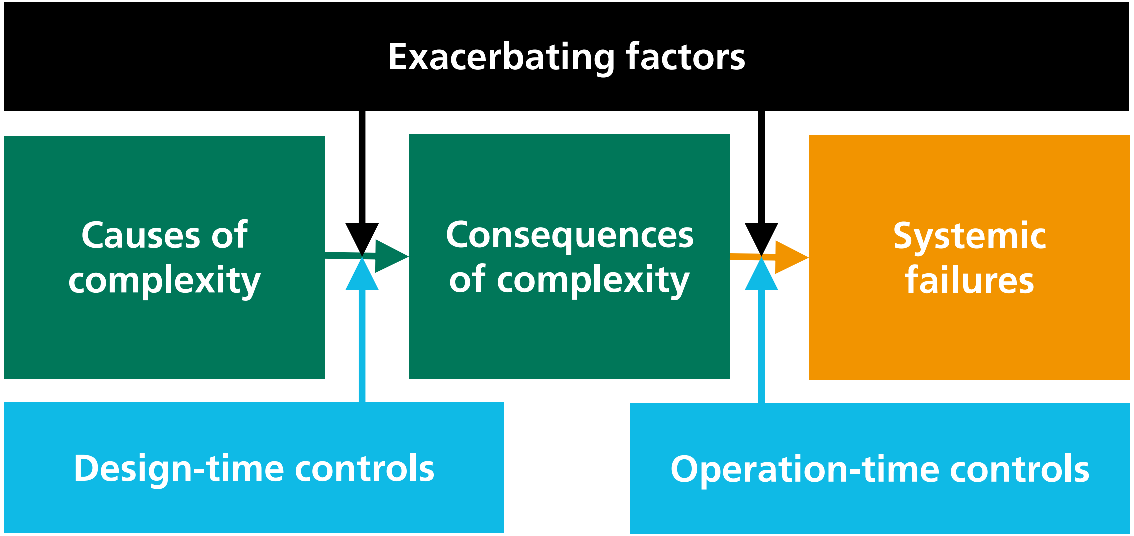
THE SAFER COMPLEX SYSTEMS FRAMEWORK

Interacting, holistic perspectives:

- Governance and Regulation
- Management and Operation
- Task, Interaction and Technical

Common factors impacting complexity and safety identified

Framework refined through analysis of 30+ Case studies



See also: https://www.raeng.org.uk/publications/reports/safer-complex-systems

UNDERSTANDING THE IMPACT OF COMPLEXITY

Uber Tempe Accident

Exacerbating factors
Politicised decision making, Casualisation of labor, …

Causes of complexity
Lag between regulation and technological change, unpredictable behaviour of environment (pedestrians), automation complacency, …

Consequences of complexity
Competing objectives, accountability gaps, semantic gaps, emergent properties, coupled feedback and inertia, …

Systemic failures
Inadequate regulation, unanticipated risks, model mismatch, decision mismatch, …

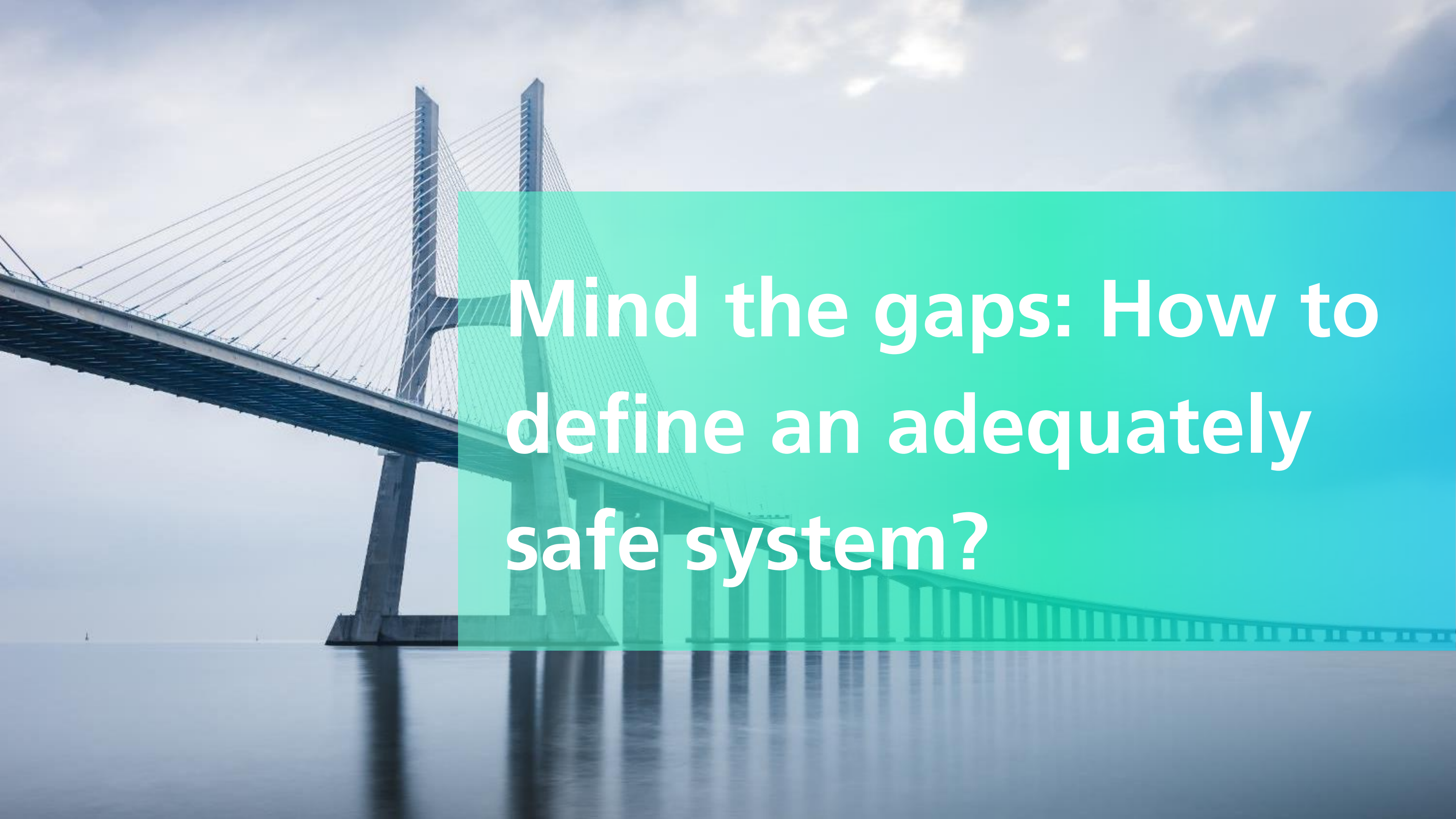
Design-time controls (ineffective)
Safety Management System, redundant technical systems, …

Operation-time controls (ineffective)
Regulatory oversight, regular test of operator capabilities, human supervision, …

Burton, S., McDermid, J. A., Garnett, P., & Weaver, R. (2021). Safety, Complexity, and Automated Driving: Holistic Perspectives on Safety Assurance. *Computer*, *54*(8), 22-32.

Research perspectives

There is a need to extend current safety management and engineering approaches to consider the **impact of complexity** and **uncertainty** within the overall system context and to establish **effective control measures**.



Mind the gaps: How to define an adequately safe system?

RECAP: SAFETY - YESTERDAY AND TOMORROW

What happens if component R213 breaks?

What impact will the system have on overall risk for a given operational domain?

SEMANTIC GAPS

Semantic Gap* – discrepancy between the intended and specific functionality, caused by:

- Complexity and unpredictability of the operational domain
- Complexity and unpredictability of the system itself
- Increasing transfer of decision function to the system

Leads to moral responsibility, legal accountability and safety assurance gaps

*Burton, Simon, et al. "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective." Artificial Intelligence 279 (2020): 103201.

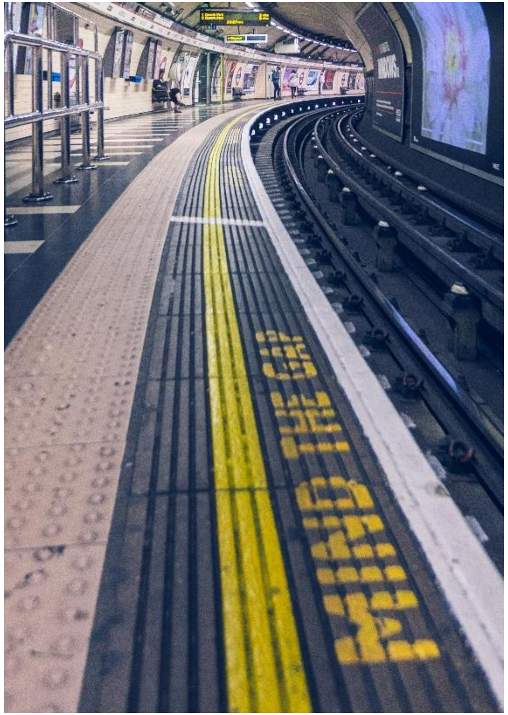


Photo by Artur Tumasjan on Unsplash

CLOSING THE GAPS

Aligned social expectations, interdisciplinary definition of desirable properties

Agile, outcome-based standards and regulations

Harmonised safety acceptance, qualification and test criteria

Resilient system designs

From simulation, to test track to open road and back again

Continuously identify and close assurance gaps, use adversarial arguments

Iteratively increasing scope of domain, interactions and system

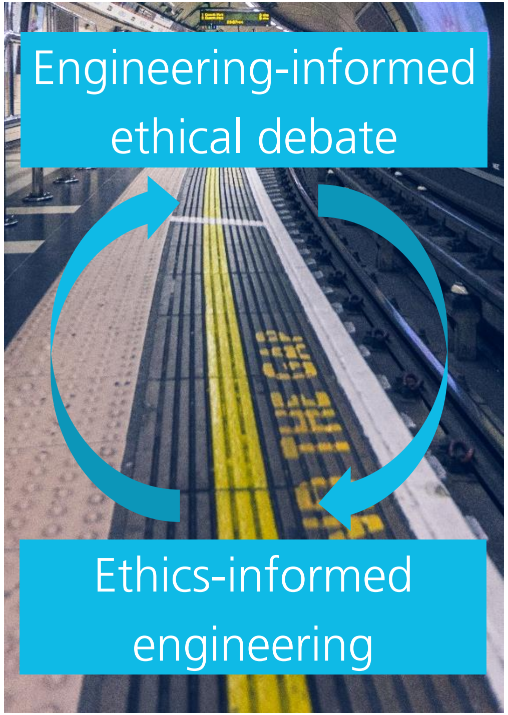


Photo by Artur Tumasjan on Unsplash

BUT HOW SAFE IS SAFE ENOUGH?
ISO TR/4804 – SAFETY AND CYBERSECURITY FOR AUTOMATED DRIVING SYSTEMS

Positive Risk Balance

Net fewer hazardous situations than human driving (e.g. collision every 300.000km)

- Definition of average human driver?
- How appropriate is the comparison with human abilities?
- What about systematic failures?
- How to measure before start of production?

Avoidance of unreasonable risk

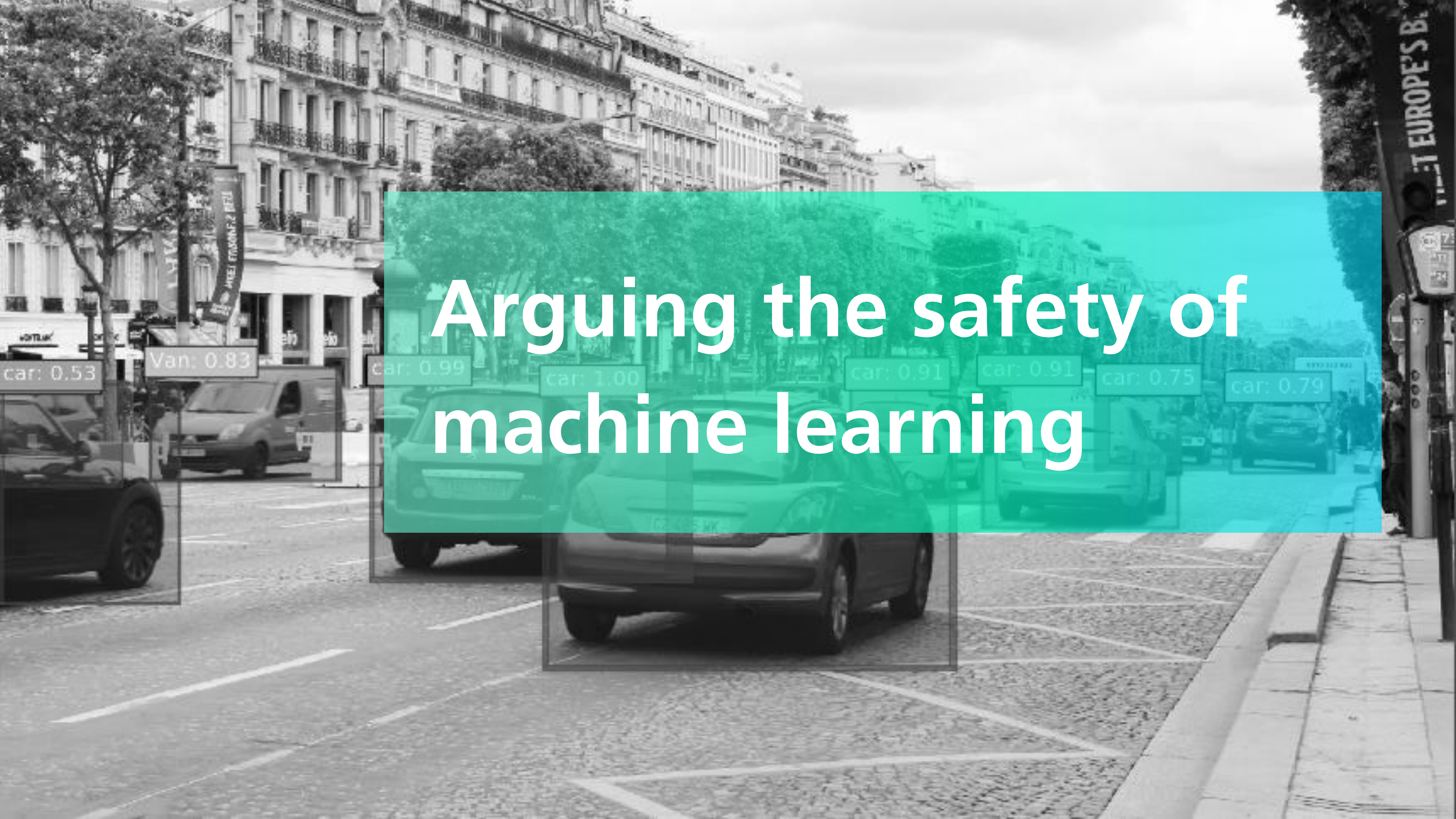
Definition of active and pro-active behaviour to avoid high-risk situations,
Application of engineering best practices and existing standards

- State-of-the-art still needs to be established
- How to define safe pro-active behaviour?
- Engineering judgement still required to determine whether system is "safe enough"

Research perspectives

A combination of various **qualitative** and **quantitative acceptance criteria** must be found that provide a convincing answer to the question: Is the system safe enough?

Arguing the safety of machine learning

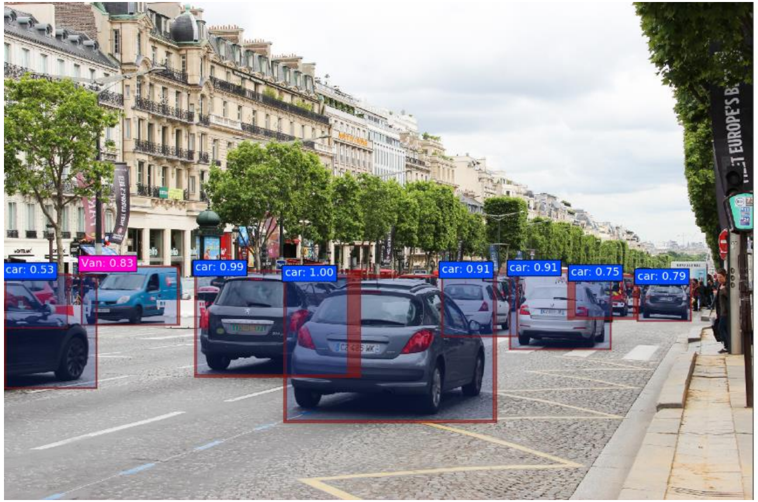


NO FREE LUNCH

Where's the catch?

- **Semantic Gap / Specification Paradox:** No explicit definition of "safe" behaviour
- **Uncertainty:** Confidence scores not necessarily indication of probability of correctness
- **Lack of explainability:** Learnt concepts are in general not understandable by humans

A safety engineer's nightmare!



„Assessing box merging strategies and uncertainty estimation methods in multimodel object detection ", Schmoeller da Roza et al., Beyond mAP: Reassessing the Evaluation of Object Detectors @ECCV 2020

BENCHMARK PERFORMANCE ≠ SAFETY

Precision: e.g. 90%

- Could mean 1/10 detected pedestrian are not really there → too many emergency stops
- Does not tell us how many pedestrians are never detected (bad)

Recall: e.g. 90%

- Could mean 1/10 pedestrians are **never** detected (bad) **or**
- For each pedestrian 1 in 10 frames are incorrect (might be o.k.)

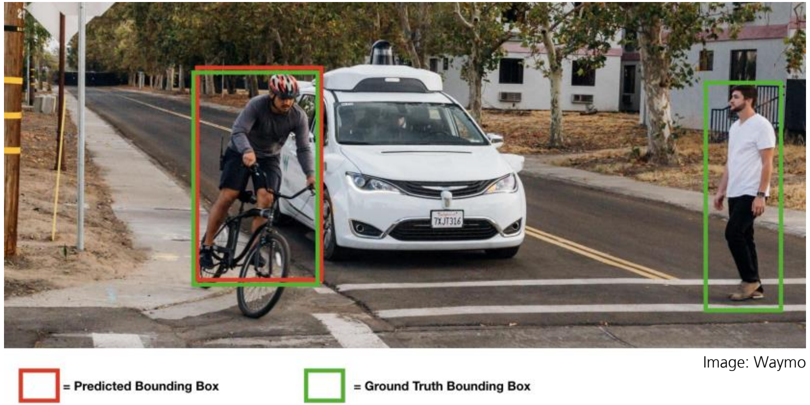


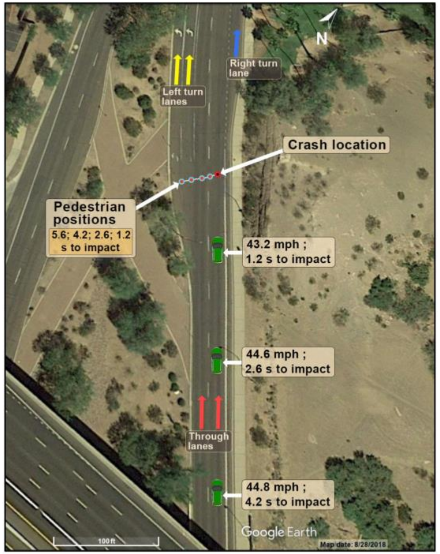
Image: Waymo

$$Precision = \frac{TP}{TP + FP} = \frac{1}{1+0} = 1$$

$$Recall = \frac{TP}{TP + FN} = \frac{1}{1+1} = 0.5$$

BENCHMARK PERFORMANCE ≠ SAFETY

- Current performance benchmarks for ML-based perception are <u>many, many </u>orders of magnitude worse than current accident rates
- Optimizing from 87% to 90% for metric X isn't going to solve the problem
- Which set of performance benchmarks could have predicted the Uber Tempe crash?
- How good should the object classifier have been?



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	<u>Classification</u>: *Vehicle*—by radar <u>Path prediction</u>: *None*; not on path of SUV	Radar makes first detection of pedestrian (classified as vehicle) and estimates speed.
-5.2	44.6	<u>Classification</u>: *Other*—by lidar <u>Path prediction</u>: *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	<u>Classification</u>: *Vehicle*—by lidar <u>Path prediction</u>: *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	<u>Classification</u>: *Vehicle*—by lidar <u>Path prediction</u>: 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	<u>Classification</u>: alternates between *vehicle* and *other*—by lidar <u>Path prediction</u>: 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	<u>Classification</u>: *Bicycle*—by lidar <u>Path prediction</u>: *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	<u>Classification</u>: *Bicycle*—by lidar <u>Path prediction</u>: 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.

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

Safety Goal:

Each pedestrian potentially within the path of the vehicle shall be safely detected

Assumptions (environment): E.g.: Size, position, movement, occlusion of pedestrians

Assumptions (system): E.g.: Image quality, capabilities of monitoring components

Acceptance criteria: Each pedestrian within the **critical range** is correctly detected with a **true positive rate** sufficient to confirm their **position** within any **sequence** of images in which the pedestrian fulfils the **assumptions**

Picture: https://www.ki-absicherung-projekt.de/

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

1

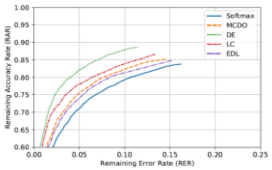
Direct measurement of failure rate of ML function

Evaluation of overall performance for a given sample space of the input domain.

Example Metrics:

Remaining Error Rate (certain but incorrect),
Remaining Accuracy rate (certain and correct),
…

Source: „Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics“, Henne et al., SafeAI 2020



Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

1

Direct measurement of failure rate of ML function

How accurate and representative are our performance predictions?

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

2

Evaluation of impact of ML insufficiencies on performance

1

Direct measurement of failure rate of ML function

How accurate and representative are our erformance predictions?

Directly measure the presence or argue the absence of specific insufficiencies.

Example Metrics:

Adversarial frequency, consistency, occlusion sensitivity, uncertainty quantification,

…

Source: „Confidence arguments for evidence of performance in machine learning for highly automated driving functions ", Burton et al., WAISE 2019

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

2

Evaluation of impact of ML insufficiencies on performance

Is there a correlation to the residual failure rate?

1

Direct measurement of failure rate of ML function

How accurate and representative are our performance predictions?

Definition of **quantitative acceptance criteria,** decomposed to ML functions

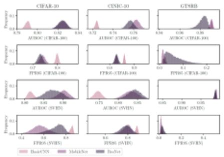
What level of performance is required?

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

How rigorously were design-time measures for reducing insufficiencies applied?

Example Metrics:

Training data selection criteria, test scenario coverage, adversarial confidence loss, ensemble diversity…



Source: „Measuring Ensemble Diversity and Its Effects on Model Robustness", Heidemann et al., AISafety 2021

1

ent

How accurate and representative are our erformance predictions?

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

3

Evaluation of effectiveness of **design-time methods** to minimise insufficiencies

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

2

Evaluation of impact of ML insufficiencies on performance

Is there a correlation to the residual failure rate?

1

Direct measurement of failure rate of ML function

How accurate and representative are our performance predictions?

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

3

Evaluation of effectiveness of **design-time methods** to minimise insufficiencies

What level of rigor in the development process is required?

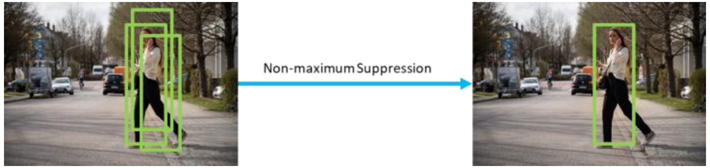
ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

Evaluation o
ML insufficie
performance

To what extent do run-time and architecture measures reduce residual failure rate?

Example Methods:

Out of distribution detection, monitors, sensor fusion…



Source: „From Black-box to White-box: Examining Confidence Calibration under different Conditions“, F. Schwaiger et al., SafeAI 2021

ur
ns?

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

3

Evaluation of effectiveness of **design-time methods** to minimise insufficiencies

What level of rigor in the development process is required?

4

Evaluation of effectiveness of **operation-time methods** to eliminate residual failures

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

2

Evaluation of impact of ML insufficiencies on performance

Is there a correlation to the residual failure rate?

1

Direct measurement of failure rate of ML function

How accurate and representative are our performance predictions?

3

Evaluation of effectiveness of **design-time methods** to minimise insufficiencies

What level of rigor in the development process is required?

4

Evaluation of effectiveness of **operation-time methods** to eliminate residual failures

What level of diagnostic coverage can be achieved?

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

ONGOING RESEARCH – HOLISTIC ARGUMENTS FOR ML SAFETY

1

Direct measurement of failure rate of ML function

How accurate and representative are our performance predictions?

2

Evaluation of impact of ML insufficiencies on performance

Is there a correlation to the residual failure rate?

3

Evaluation of effectiveness of **design-time methods** to minimise insufficiencies

What level of rigor in the development process is required?

4

Evaluation of effectiveness of **operation-time methods** to eliminate residual failures

What level of diagnostic coverage can be achieved?

Definition of **quantitative acceptance criteria,** decomposed to ML functions

What level of performance is required?

Research perspectives

Derive a set of ***meaningful safety metrics and methods for AI***, apply within a ***holistic and iterative approach*** to building an argument for safety within a specific environmental and system context.

STANDARDISATION - NEXT STEPS
ISO/PAS 8800 ROAD VEHICLES – SAFETY AND AI

Context

Currently fragmented and incomplete standards w.r.t. AI and safety for automotive applications

Generic standard ISO/TR 5469 Functional safety and AI systems under development

ISO PAS 8800

New Publicly Available Specification to provide guidance on applying automotive safety standards to AI-based functions

Status

ISO/TC 22/SC32/WG 14 founded, Kick-off 2021-12-07

Project lead: S. Burton, Fraunhofer IKS

ISO/TC 22 **N 4142**

ISO/TC 22 "Road vehicles"
Secretariat: **AFNOR**
Committee Manager: **Maupin Valérie Mme**

N4142_NWIP on ISO PAS 8800_For ballot before 2021-09-01 (SC32)

Document type	Related content	Document date	Expected action
Ballot / Reference document		2021-06-30	VOTE by 2021-09-01

Road Vehicles – Safety and Artificial Intelligence

1 Scope

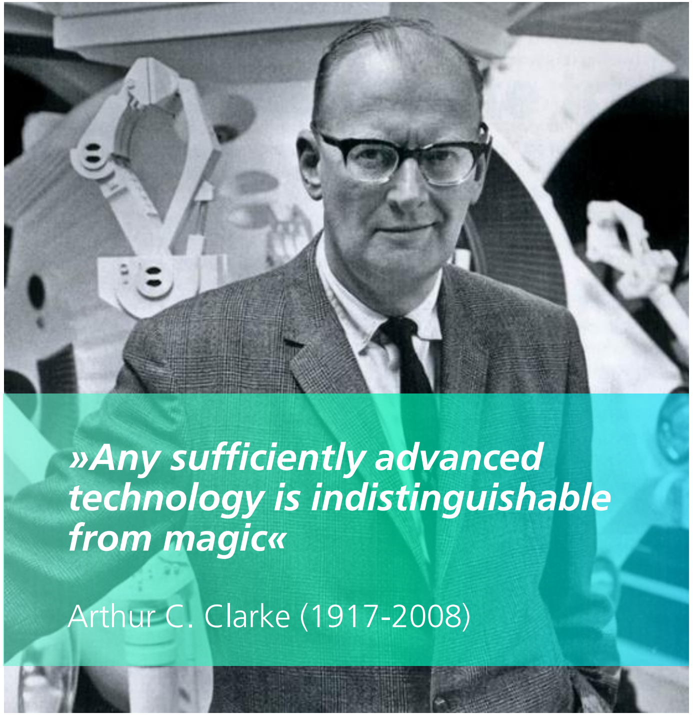
This document defines safety-related properties and risk factors impacting insufficient performance and malfunctioning behaviour of Artificial Intelligence (AI) within a road vehicle context. It describes a framework that addresses all phases of the development and deployment lifecycle. This includes the derivation of suitable safety requirements on the function, considerations related to data quality and completeness, architectural measures for the control and mitigation of failures, tools used to support AI, verification and validation techniques as well as the evidence required to support an assurance argument for the overall safety of the system.

Conclusions

RESEARCH DIRECTIONS FOR AI SAFETY AND AUTONOMOUS SYSTEMS

Acknowledge system complexity:

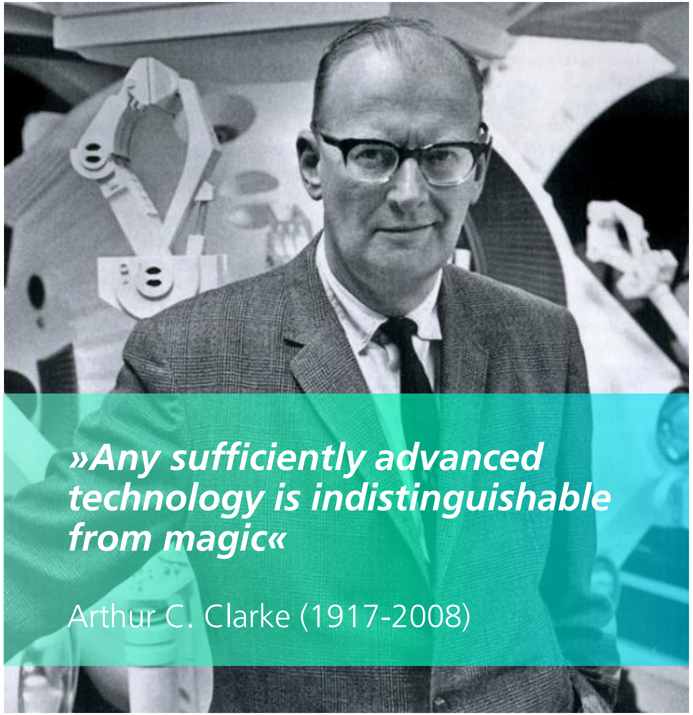
- Address the problem not only from a technical perspective
- by engaging in an **engineering-informed interdisciplinary dialog**
- and applying **ethically-informed engineering practices**



RESEARCH DIRECTIONS FOR AI SAFETY AND AUTONOMOUS SYSTEMS

Acknowledge system complexity:

— Apply **systems engineering approaches** that use an optimal combination of domain understanding as well system design, verification and validation measures to mitigate risk



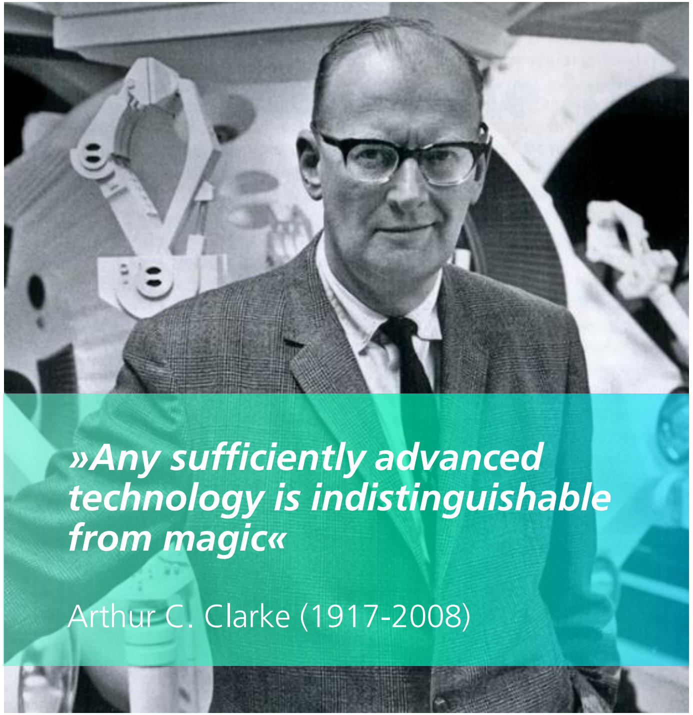
RESEARCH DIRECTIONS FOR AI SAFETY AND AUTONOMOUS SYSTEMS

Take the magic out of AI:

— **Which levels of performance** are actually required of the machine learning function?

— Can an **acceptable level of performance** ever be met?

— How **effective** are different methods of collecting evidence?



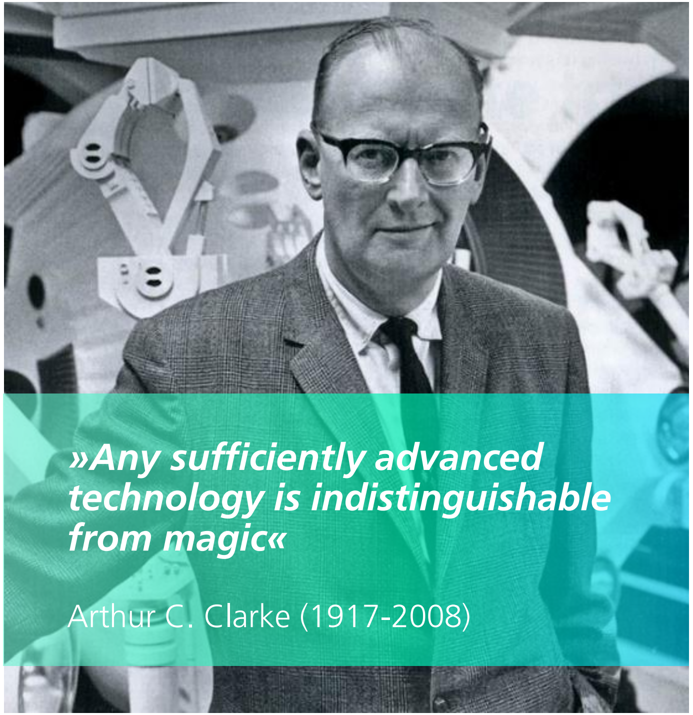
RESEARCH DIRECTIONS FOR AI SAFETY AND AUTONOMOUS SYSTEMS

Take the magic out of AI:

This requires…

— A **structured, iterative process** for ensuring a systematic application of appropriate methods during development is required

— …and a fundamental understanding of the **limitations of AI-methods** and a formalisation of safety-relevant measurements and metrics



Thank you for your attention

Any questions?