Fraunhofer IKS

Fraunhofer Institute for Cognitive Systems IKS

2022-06-01 GfSE e.V. und Systems Engineering

Prof. Dr. Simon Burton

Safety assurance under uncertainty

Cognitive Safety-Critical Cyber Physical Systems

Cognitive systems are software-intensive technical systems that imitate cognitive capabilities such as perception, learning, and reasoning.

Automated driving

Industrial Robotics

Driverless trains

Medical devices

Traditional Approach to Safety

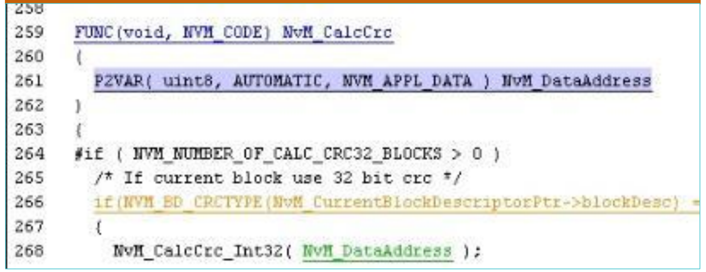
Functional safety (ISO 26262):
Absence of unreasonable risk due to hazards caused by **malfunctioning behaviour** of E/E systems

Risk associated with malfunctioning behaviour

Random hardware errors

Photo: Christian Taube - Own work

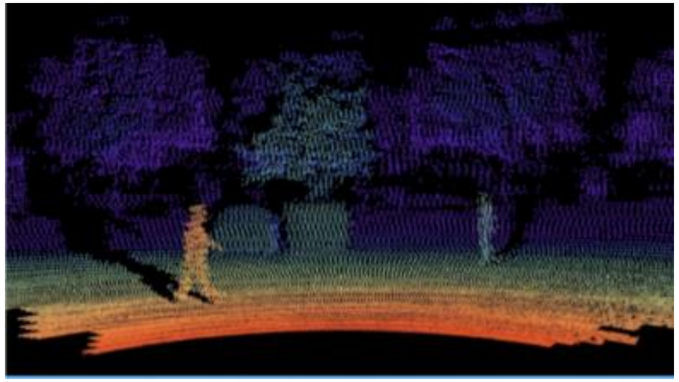
Systematic errors (HW and SW)



Picture: Mathworks

What‘s changing?

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



Source: https://velodynelidar.com

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

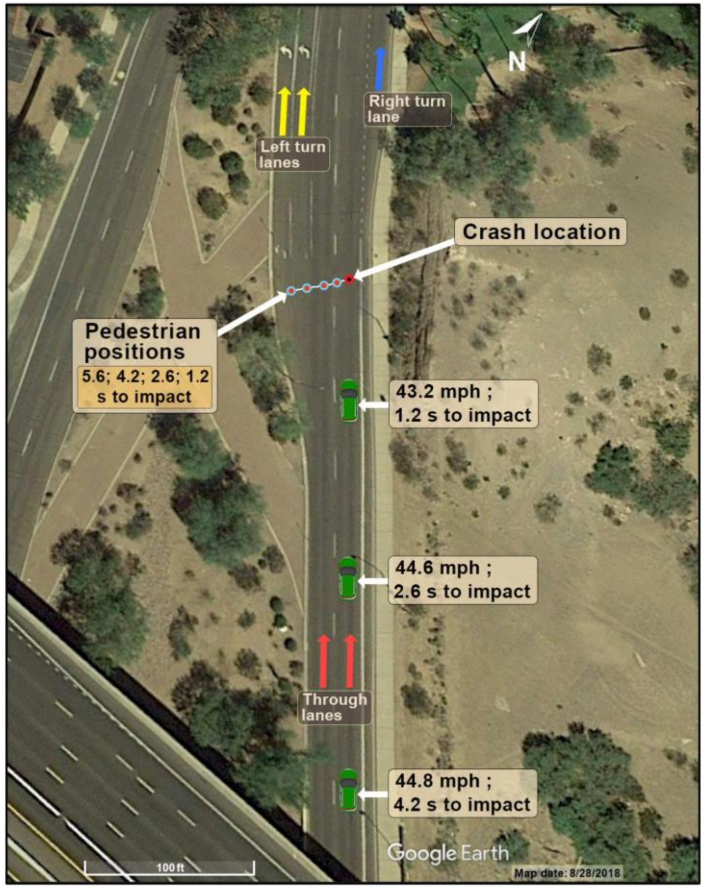
Scope & unpredictability of operational domain and critical events

Inaccuracies & noise in environmental sensors and signal processing

Heuristics or machine learning techniques with unpredictable results

Case Study – Uber Tempe incident

Interacting layers of complexity and uncertainty



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.
-4.2	44.8	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.
-3.9[c]	44.8	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.
-3.8 to -2.7	44.7	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.

ISO 21448: Failure of the intended functionality: *Absence of unreasonable risk due to hazards resulting from* ***functional insufficiencies*** *of the intended functionality or by reasonably* ***foreseeable misuse*** *by road users*

Triggering condition: *Specific conditions of a scenario that serve as an initiator for a subsequence system reaction contributing to either a hazardous behaviour or an inability to prevent or detect and mitigate a reasonably foreseeable indirect misuse*

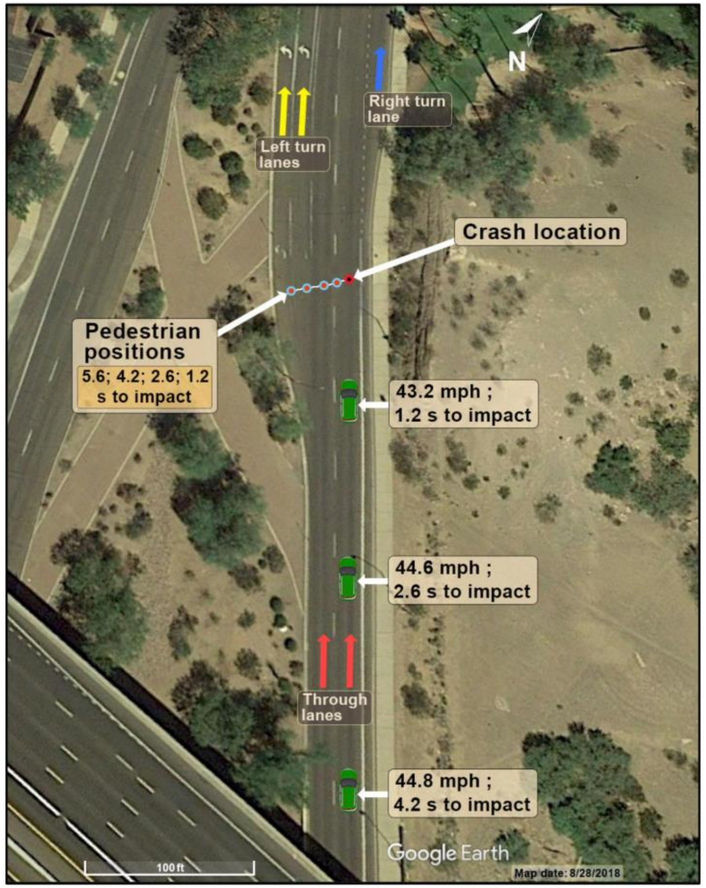
Technical

Failure of system to correctly detect pedestrian and avoid collision

Fraunhofer IKS

Case Study – Uber Tempe incident

Interacting layers of complexity and uncertainty



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.
-4.2	44.8	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.
-3.9[c]	44.8	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.
-3.8 to -2.7	44.7	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.

ISO 21448: Failure of the intended functionality: *Absence of unreasonable risk due to hazards resulting from **functional insufficiencies** of the intended functionality or by reasonably **foreseeable misuse** by road users*

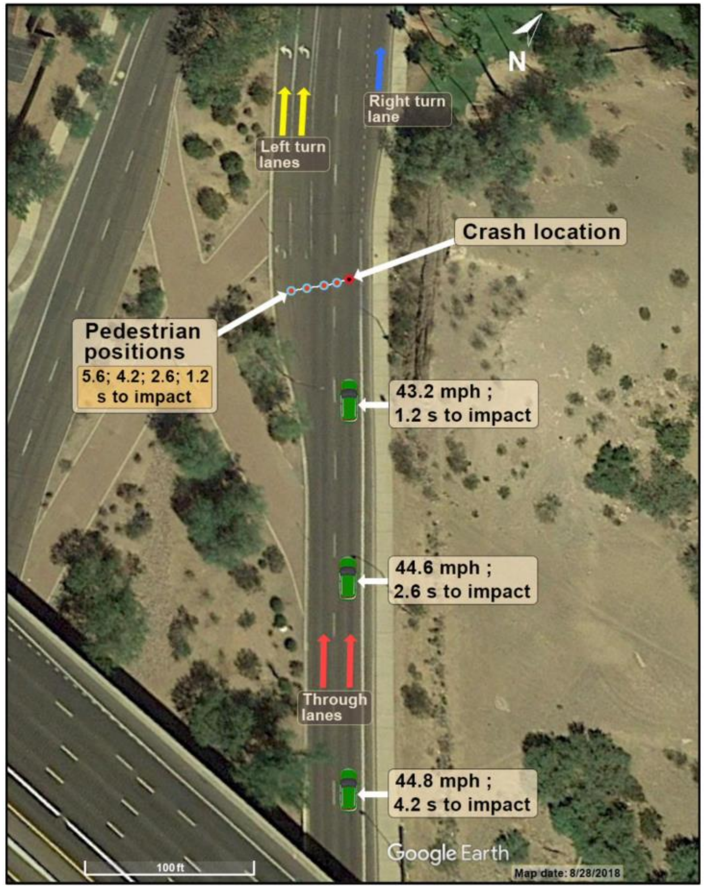
Indirect misuse: *E.g. lack of monitoring by the human operator due to Automation Complacency*

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

Case Study – Uber Tempe incident

Interacting layers of complexity and uncertainty



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.
-4.2	44.8	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.
-3.9[c]	44.8	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.
-3.8 to -2.7	44.7	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.

Governance — Failure to regulate accountability for safety of automated driving

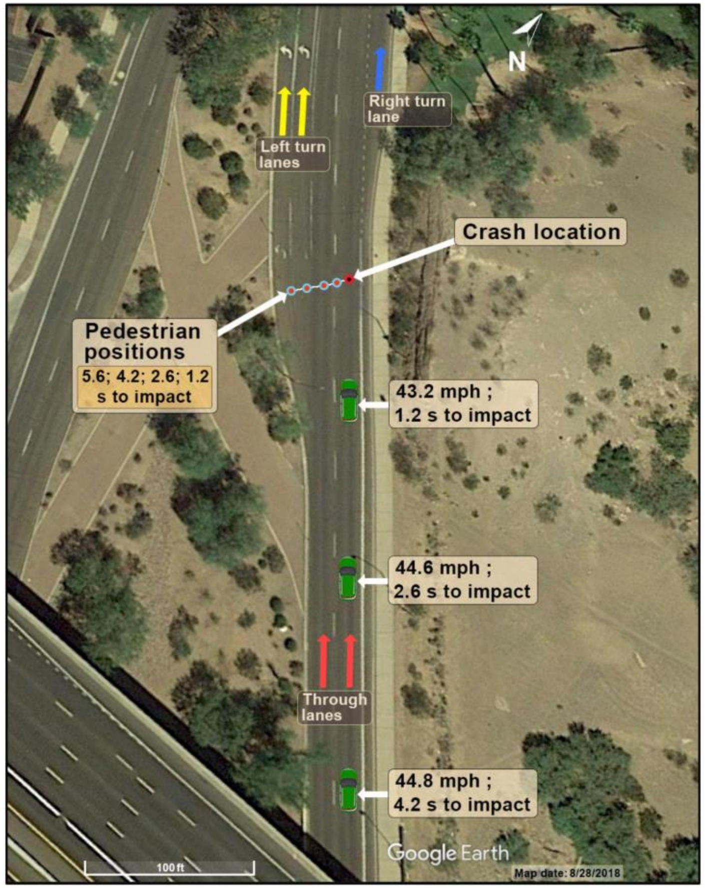
Management — Inadequate engineering and operating processes, lack of oversight of safety driver

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

Systemic Failures

Consequences of system complexity and uncertainty



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

Systemic Failures*:

Failure at a system level caused by interactions between behaviours of the systems components and interactions with or dependencies with its environment, e.g.:

- **Governance:** Inadequate deployment decisions, inadequate regulatory control
- **Management and operations:** Accountability mismatch, unanticipated risks
- **Human factors and technical:** model mismatch, decision mismatch, authority mismatch

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

Fraunhofer IKS

Safety is becoming less about what happens when individual technical components break and more about managing the emergent risk associated with increasing complexity

System complexity

Emergent properties of complex systems

A **complex system** exhibits behaviours that are **emergent properties** of the interactions between the parts of the system, where the behaviours would **not be predicted** based on knowledge of the parts and their interactions alone.

Caused by:

- Semi-permeable boundaries
- Non-linearity, mode transitions, tipping points
- Self-organization and ad-hoc systems

Consequences of system complexity

The semantic gap

Semantic Gap* – discrepancy between the intended and specified 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

*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.

Uncertainty:

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

*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.

Fraunhofer IKS

Uncertainty

Dimensions of uncertainty

Location:

Environment uncertainty includes uncertainty in the execution context of the system and uncertainty arising from the unpredictable behavior of humans that the system interacts with.

Goals uncertainty manifests due to imprecise specification, modeling, and derivation of the system's goals (including safety goals).

Model uncertainty results from the failure to adequately model the system, its environment, or the behavior thereof.

Functions uncertainty is caused by system functions with non-deterministic, inaccurate, or unexpected effects and side-effects.

Resources uncertainty occurs because of changes to the essential components of the system, e.g. due to failures, resulting in some of the system's functionality or resources becoming unavailable.

Levels of uncertainty:

Statistical uncertainty can be expressed in statistical terms, such as with probability distributions or using belief theory (Quantitative).

Scenario uncertainty can only be described using scenarios, which are plausible states of the system and/or its environment without any statistical support (Qualitative).

Lack of awareness means the system is not aware that its knowledge is subject to uncertainty ➔ **Requires measures external to the system**

Uncertainty

Technical uncertainty and machine learning

Data as the specification:

- No explicit definition of "safe" behaviour

Complex operational design domain

- Distributional shift / scalable oversight: Dealing with rare but critical events and changes in the environment over time

Robustness and generalisation:

- (Correct) outputs sensitive to small changes in the inputs

Prediction uncertainty:

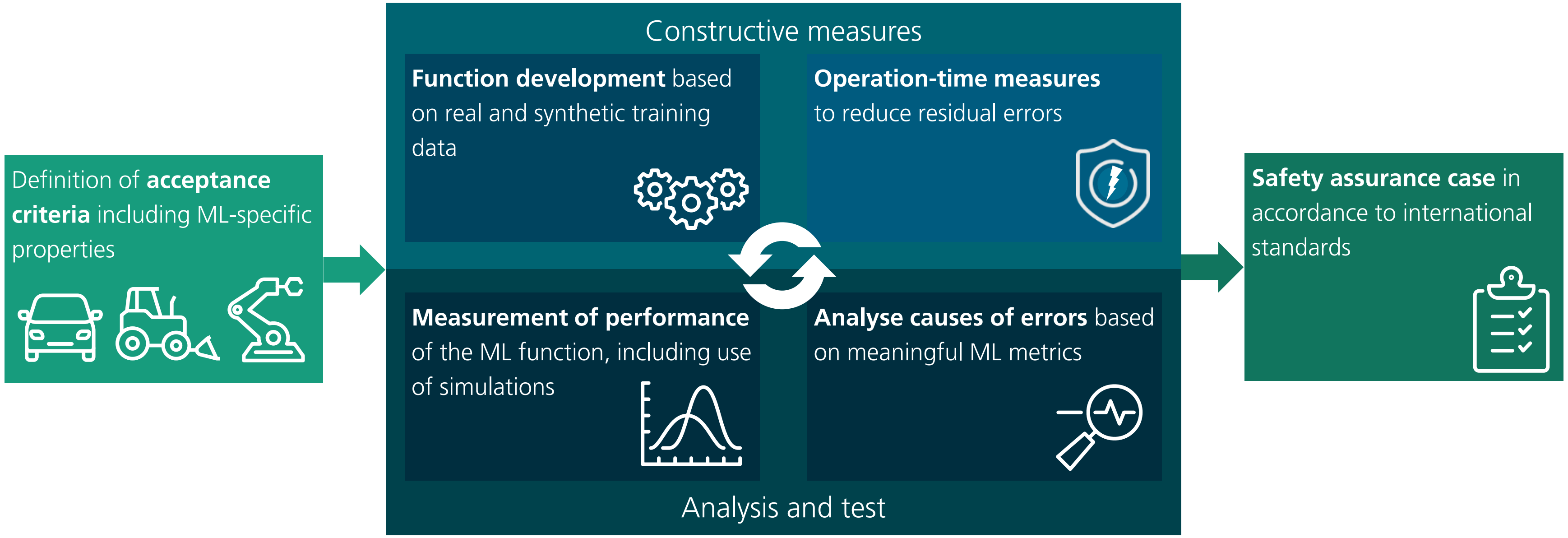
- Confidence scores not necessarily indication of probability of correctness

Explainability:

- Learnt concepts are not understandable by humans

Uncertainty

Reducing uncertainty in machine learning: Safe-ML-Ops



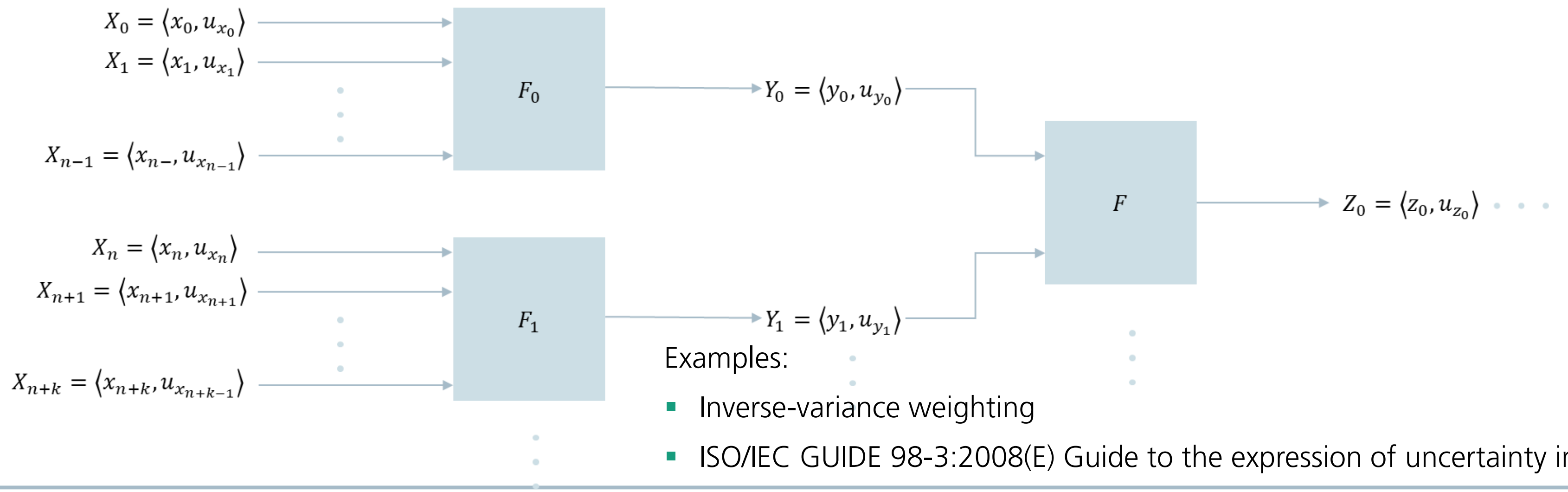
- Iterative development based on increasing understanding of the performance characteristics of the ML function, influence of environmental factors and the effectiveness of measures to reduce the impact of residual errors
- Depends on a causal understanding of the sources of uncertainty and errors in the ML function
- **Objective:** realistic evaluation of the **statistical and scenario uncertainty** in the ML function

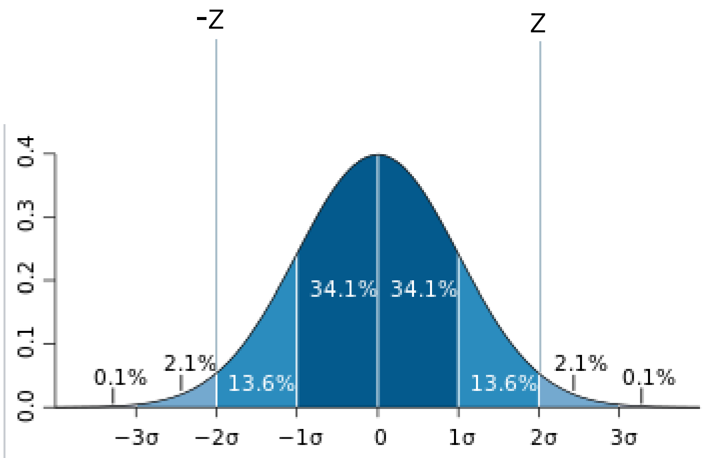
Uncertainty

Understanding the technical impact of uncertainty within the system

Uncertainty quantification and propagation @ runtime as potential measure to:

- **Analysis steps:** Uncertainty estimation, design of architectural uncertainty mitigation patterns, uncertainty propagation
- Can relax worst-case assumptions through risk-awareness of current context and thus increase the system's utility
- Only applicable to addressing quantifiable **statistical uncertainty** (e.g. can be represented by a Gaussian distribution)



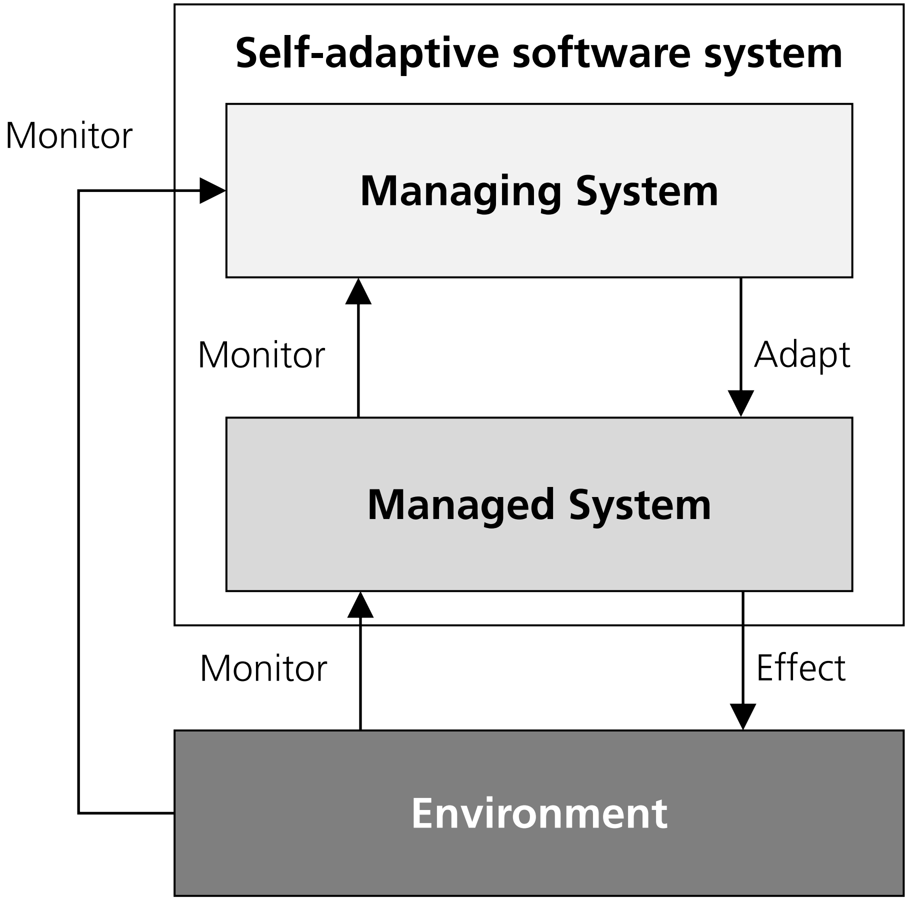


Examples:

- Inverse-variance weighting
- ISO/IEC GUIDE 98-3:2008(E) Guide to the expression of uncertainty in measurement

Uncertainty

Self-adaptive systems



Self-adaptive systems:

Provide resilience against faults and uncertainties within the system as well as uncertainties and changes within environment

Assurance challenges*:

Perpetual assurance: continuous generation of evidence that system requirements are met, despite adaptation of system and environment

Composing assurances: avoiding re-validation for emergent systems (-of-systems)

Feedback and monitoring: defining observation points for determining when the assurance process is not effective

*Lemos, Rogério de, et al. "Software engineering for self-adaptive systems: Research challenges in the provision of assurances." *Software Engineering for Self-Adaptive Systems III. Assurances*. Springer, Cham, 2017. 3-30.

Uncertainty

Types of uncertainty and mitigation techniques

Specification uncertainty
- Completeness
- Implicit assumptions
- Competing objectives
- Test coverage
- …

Technical uncertainty
- Sensing insufficiencies
- ML for perception, planning
- Actuating inaccuracies
- Security vulnerabilities
- …

Assurance uncertainty
- Completeness
- Validity of evidence
- Stability over time
- Monotonic safety
- Statistical confidence
- …

Design-time controls
- Standardised and restricted domain ontologies
- Field validation in silent mode
- Uncertainty quantification/propagation analysis
- Qualitative evaluation of assurance case confidence
- …

Operation-time controls
- Technical redundancy and monitoring
- Run-time uncertainty quantification
- Self-adaptation & dynamic risk management
- Dynamic assurance cases
- …

Uncertainty

Regulation and assurance gaps

SAFETY ENGINEERING

Clearly defined and measurable criteria for Trustworthy AI

GAPS

Regulations for AI

Trustworthy AI*:

- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability
- Human agency and oversight

SOCIETAL EXPECTATIONS

Does the system fulfill all the technical criteria required to be considered trustworthy?

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

*Source: Ethics Guidelines for trustworthy AI, Independent high-level expert group on Artificial Intelligence, EU Commission, 2019

Moving towards complexity and uncertainty aware safety assurance

Safety assurance under uncertainty
Summary

Each iteration of technologies introduces new challenges to safety assurance, currently:

- Increasing automation within an open context
- Use of AI/Machine Learning for safety-critical functions

Systems engineering and safety assurance methodologies need to adapt to these challenges

Safety arguments are only as strong as the confidence in the information they rely on

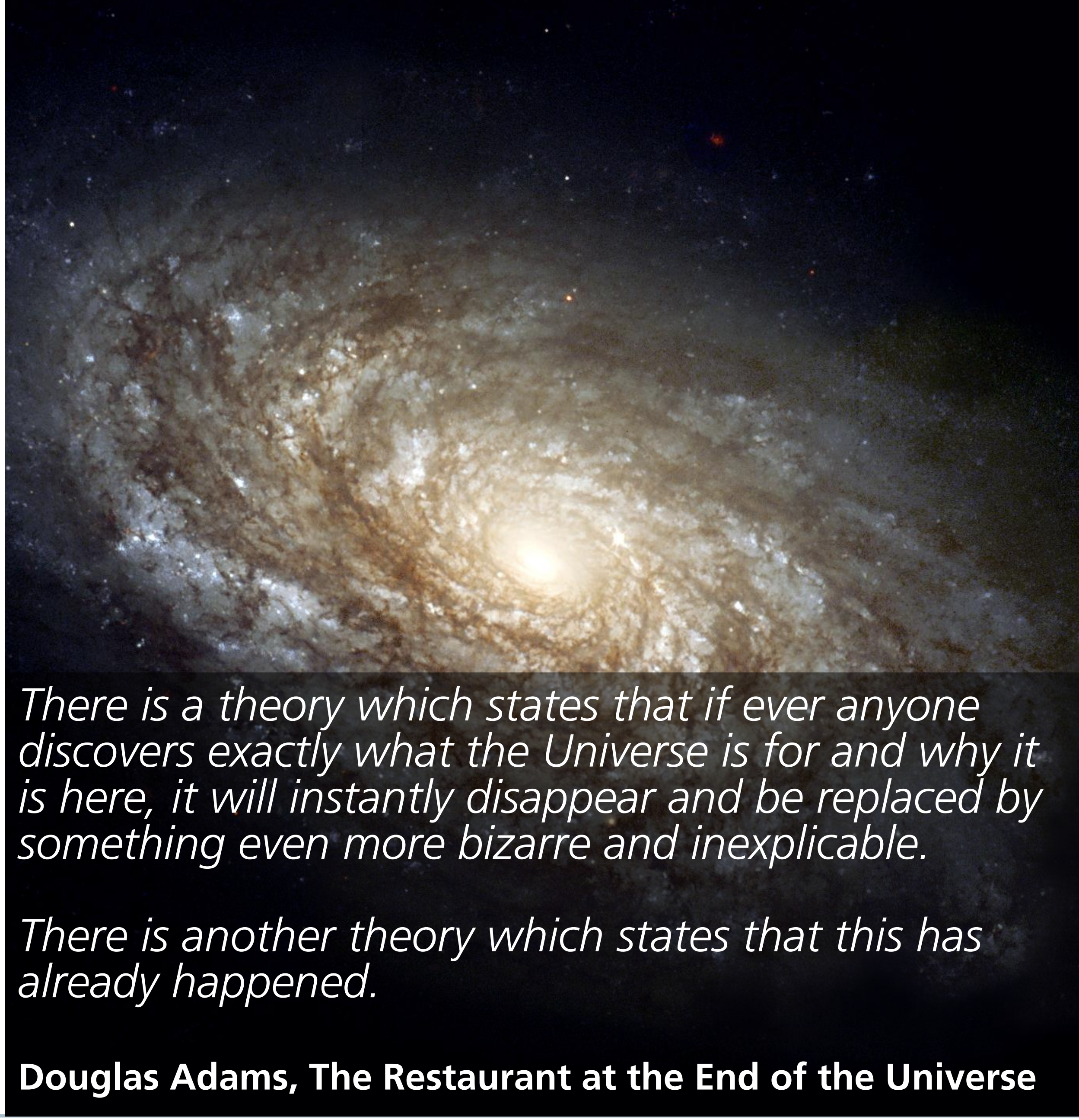


Image: NASA

  Public information

Fraunhofer IKS

Safety assurance under uncertainty

Summary

The topic of uncertainty in safety-critical systems is currently covered from a number of disparate perspectives. There is a need for:

- Common definitions of various types of uncertainty impacting the safety of highly automated AI-based cyber-physical systems
- Overarching development and assurance methodology

This is an evolving discipline, see similar "calls for action":

Calinescu, Radu, et al. "Understanding uncertainty in self-adaptive systems." 2020 IEEE international conference on autonomic computing and self-organizing systems (acsos). IEEE, 2020.

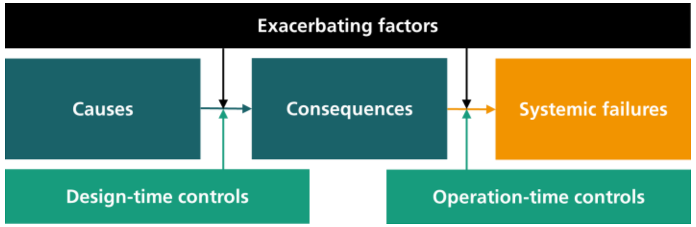
Harel, David, Assaf Marron, and Joseph Sifakis. "Autonomics: In search of a foundation for next-generation autonomous systems." Proceedings of the National Academy of Sciences117.30 (2020): 17491-17498.

Safety assurance under uncertainty

Complexity-aware systems safety engineering

Understand complexity and sources of uncertainty:

Regulatory factors,
System operation and management,
Human-Machine interaction,
Technical uncertainty



Holistic Domain and Requirements Analysis

Legal, societal and ethical expectations, ODD standards, System theoretic safety analyses,…

Resilient System Architectures

Uncertainty quantification, Self-adaptive systems, dynamic risk management,…

Continuous Verification and Validation

Simulation, extreme value analysis, field testing,…

Continuous Assurance

that a tolerable residual level of risk has been achieved, is maintained and residual assurance gaps closed in the field

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

  Public information

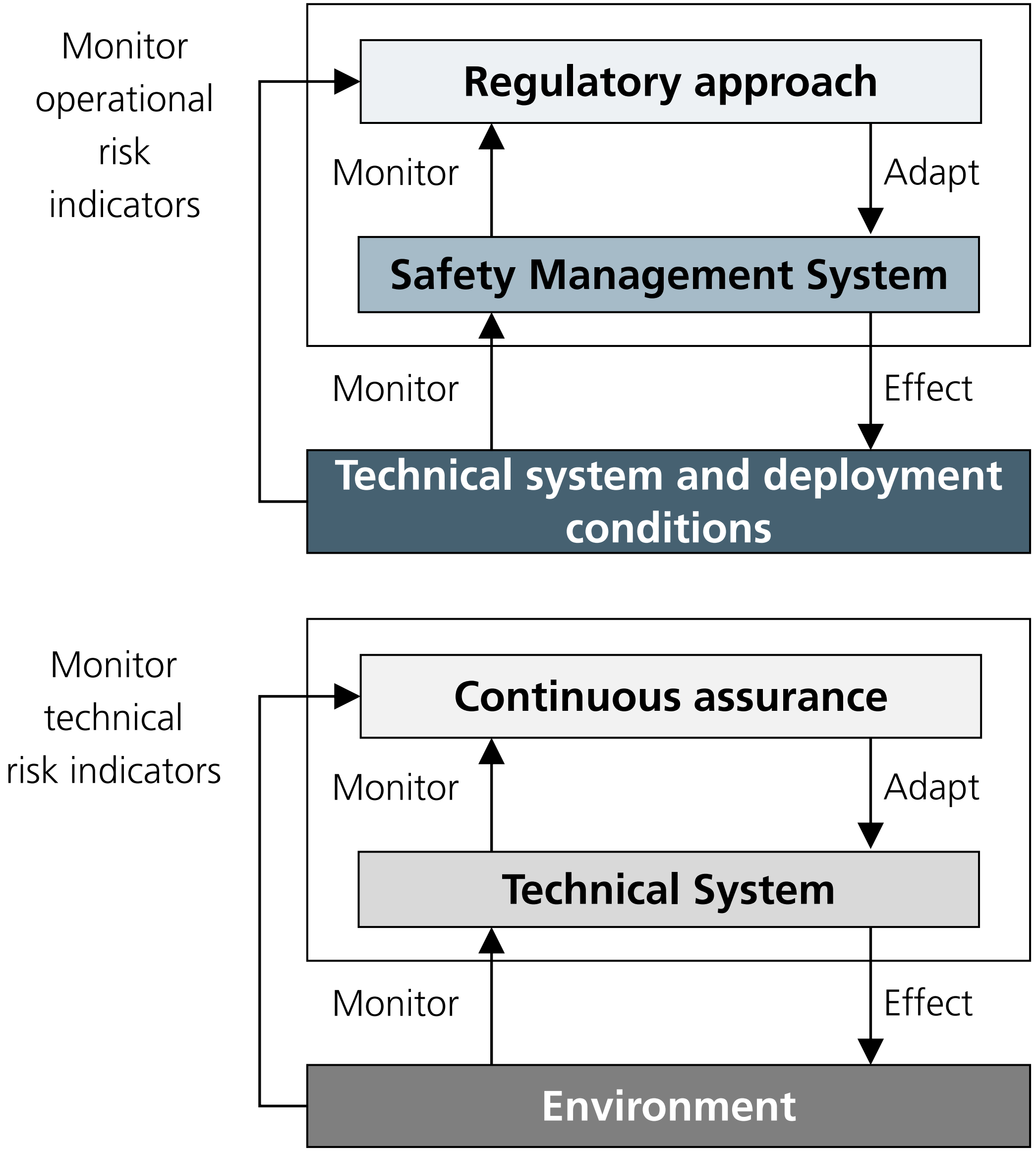
Fraunhofer IKS

Safety assurance under uncertainty

Pragmatic next steps

Deliberate and planned bootstrapping approaches should be taken to increasing operational context and functional scope, whilst monitoring impact of complexity and uncertainty

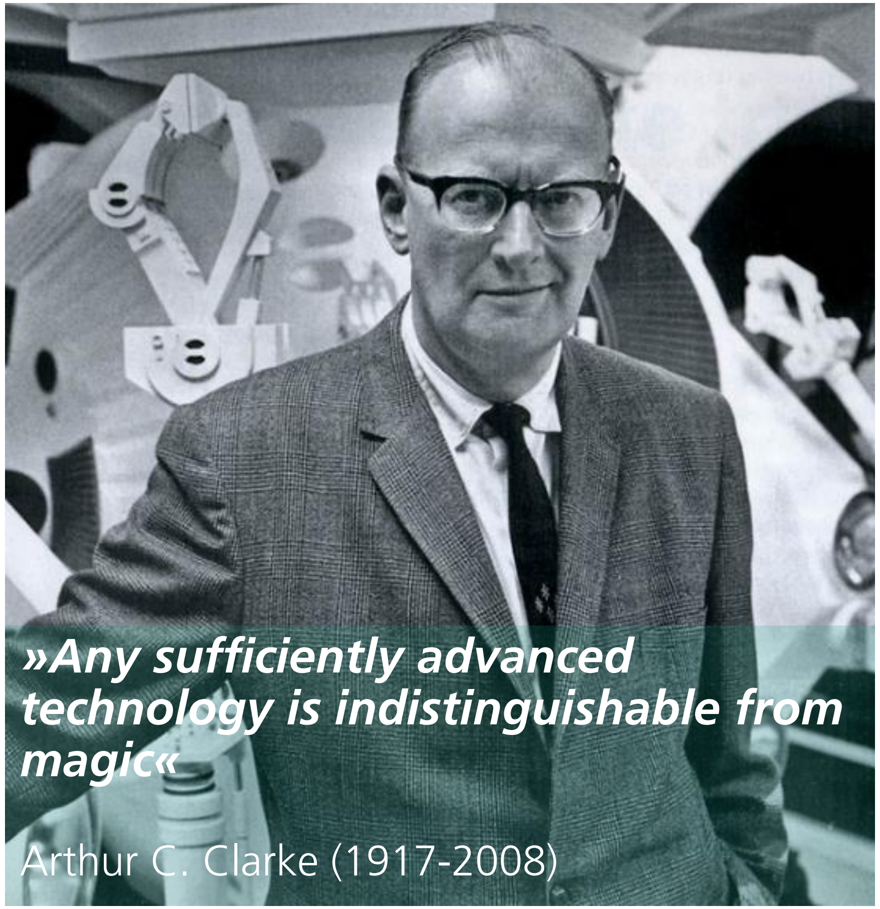
- Requires a calibrated level of tolerable residual risk
- Observation points must be defined to act as early warning indicators for increased risk/uncertainty
- Should be considered with the phased introduction of regulation and standards for automated driving (ALKS, Highway Chauffer, Delivery Drones,…)
- Applied to across all layers of governance, management and operations, human factors and technical systems



Safety assurance under uncertainty

Some ongoing research questions

- Definition of risk acceptance criteria for complex highly-automated systems
- Bridging the gap between societal and ethical expectations and technical acceptance criteria
- Role of quantitative and qualitative evidence in assuring the safety of highly automated AI-based cyber-physical systems under uncertainty
- Safety assurance of AI/ML
- Uncertainty propagation analysis during design and run-time uncertainty quantification
- Safety assurance of self-adaptive systems



Safety assurance under uncertainty
Wrap up

- **Quantitative arguments** alone are not feasible due to uncertainty in setting targets as well as in demonstrating that they are met
- **System level -1 view** required to evaluate the context of the systems and determine causes and impact of emergent complexity and uncertainty
- The assurance process must **acknowledge causes and consequences of complexity and uncertainty**
- Iterative **"Safe Dev Ops"** approaches are inevitably required in order continuously uncover and minimize residual uncertainties in the system and assurance case
- **Successive scenario-based validation and deployment** nevertheless recommended to limit scope and allow for a targeted evaluation of triggering conditions

To understand the path to safe, highly automated AI-based cyber-physical systems, it is essential to acknowledge sources of uncertainty within the safety assurance process.

A holistic view of the system and its environment is required during design and operation to manage the emergent risk of ever more complex systems.

Contact

Prof. Simon Burton
Research Division Director, Safety Assurance
Tel. +49 89 547088-341
simon.burton@iks.fraunhofer.de

Fraunhofer IKS
Hansastrasse 32
80686 München
www.iks.fraunhofer.de