Fraunhofer IKS

Fraunhofer Institute for Cognitive Systems IKS

TECoSA Seminar – 2023-04-06

Prof. Dr. Simon Burton

Addressing Uncertainty in the Safety Assurance of Machine Learning

Overview of the presentation

This presentation is based on the recently published article:
Burton S., Herd. B. Addressing uncertainty in the safety assurance of machine-learning" Frontiers of computer science, Vol 5., 2023: 10.3389/fcomp.2023.1132580, https://www.frontiersin.org/articles/10.3389/fcomp.2023.1132580

SAFE INTELLIGENT SYSTEMS

Safety & Trustworthiness

Absence of unacceptable *risk* of harm to persons or the environment

Demonstrably dependable: utility, reliability, availability,…*

**Can also include properties such as cost, privacy, security, etc. We require systems that are safe and yet still able to provide the required functional value (utility), under specific constraints (e.g. cost)*
May also require the alignment between technical capabilities and ethical expectations

Cognitive cyber-physical systems

Achieve higher levels of automation by implementing or mimicking cognitive abilities such as perception, reasoning, learning and adaptation

Integrate sensing, computation, control and networking into physical objects and infrastructure, connecting them to the internet and each other

Traditional approach to safety

Functional Safety:

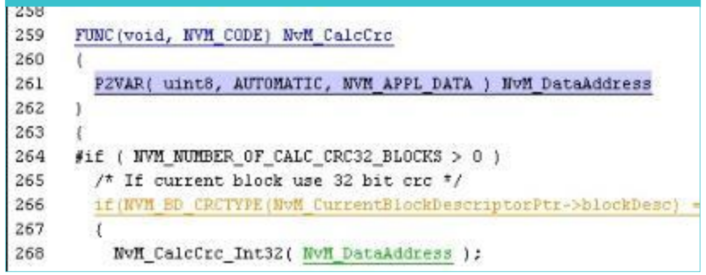
"Absence of unreasonable **risk** due to hazards caused by **malfunctioning behaviour** of E/E systems"

Risk associated with malfunctioning behaviour

Random hardware errors

Systematic errors (HW and SW)

Photo: Christian Taube - Own work



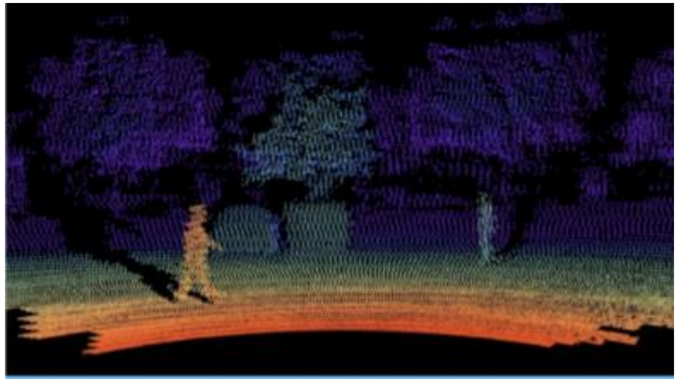
```
FUNC(void, NVM_CODE) NvM_CalcCrc
(
  P2VAR( uint8, AUTOMATIC, NVM_APPL_DATA ) NvM_DataAddress
)
{
#if ( NVM_NUMBER_OF_CALC_CRC32_BLOCKS > 0 )
  /* If current block use 32 bit crc */
  if(NVM_BD_CRCTYPE(NvM_CurrentBlockDescriptorPtr->blockDesc) =
  {
    NvM_CalcCrc_Int32( NvM_DataAddress );
```

Picture: Mathworks

What's changing?

Increasing complexity and uncertainty in cognitive cyber-physical systems

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

Overview of the presentation

Definitions of complexity and uncertainty

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 e.g.:

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

See: 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.

  Public information

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

Definitions of complexity and uncertainty

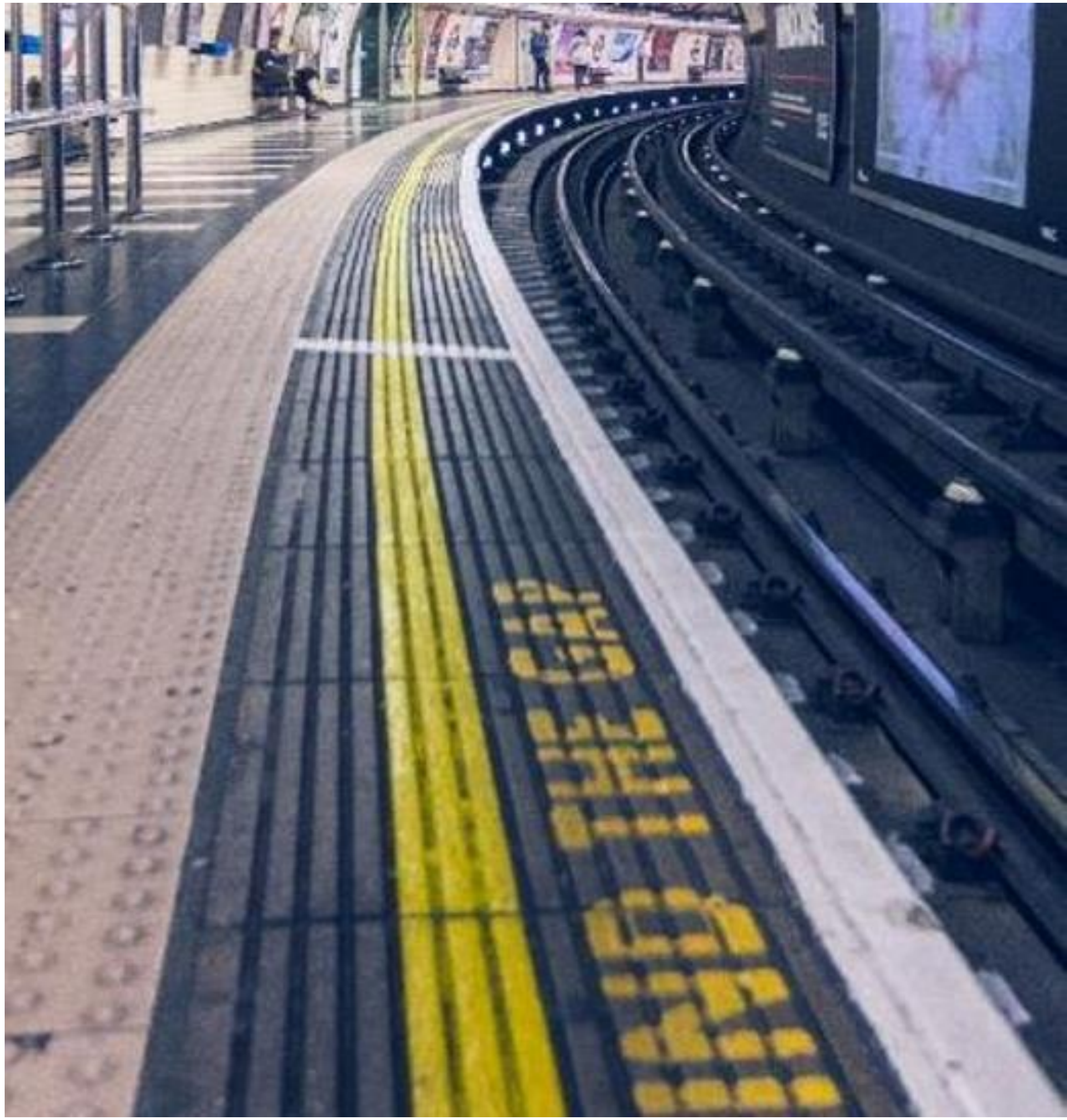
Complexity of cognitive cyber-physical systems

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

Lead to *Semantic Gaps – discrepancy between the intended and specified functionality.***

Leads to hazardous systemic failures, moral responsibility gaps, liability gaps and ***safety assurance gaps***

*Burton, Habli, Lawton, McDermid, Morgan, Porter. "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective." Artificial Intelligence 279 (2020): 103201.



Definitions of complexity and uncertainty

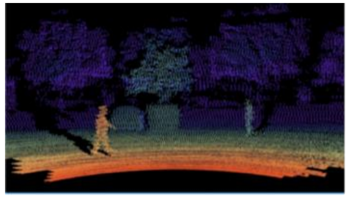
Definitions of uncertainty

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

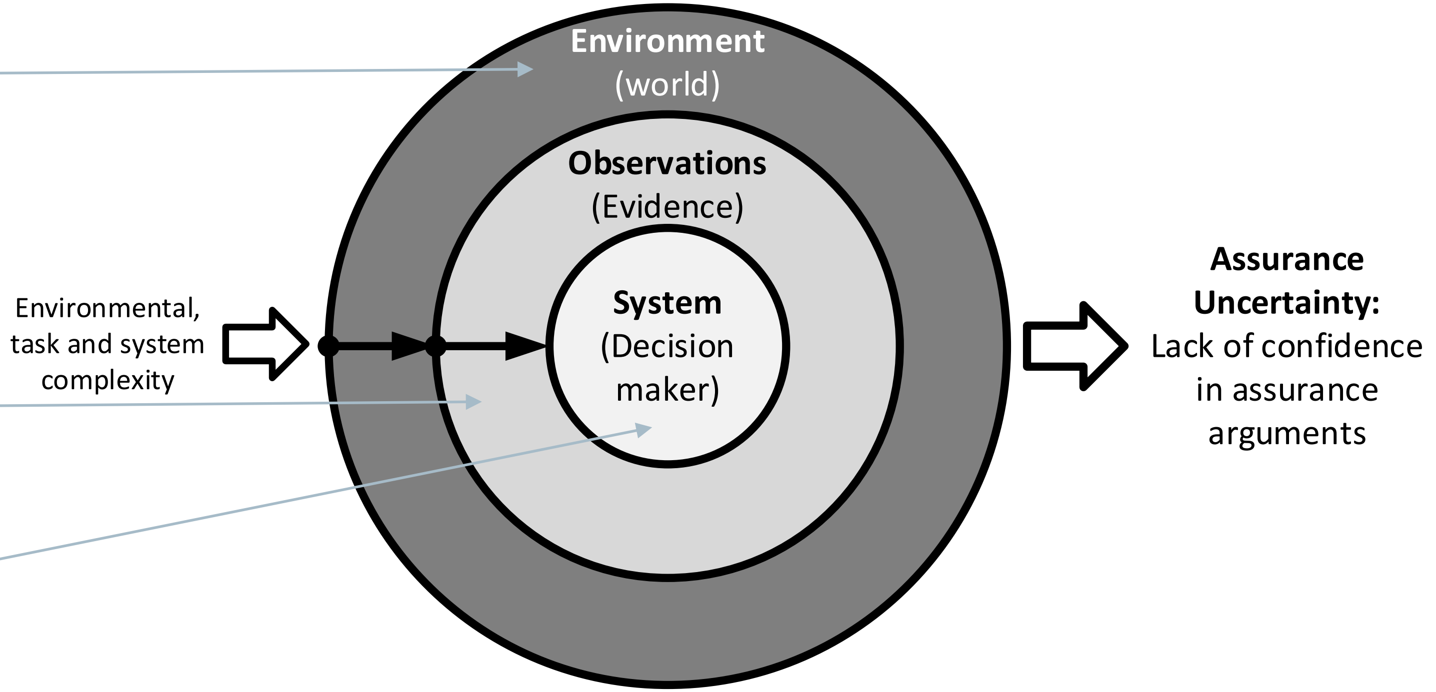


Source: https://velodynelidar.com

Inaccuracies & noise in environmental sensors and signal processing

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

Heuristics or machine learning techniques with unpredictable results



Manifestations of uncertainty

Interpretation of: Lovell, B. E. (1995). A taxonomy of types of uncertainty

*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.

Definitions of complexity and uncertainty

Relative definitions of uncertainty

Uncertainty:

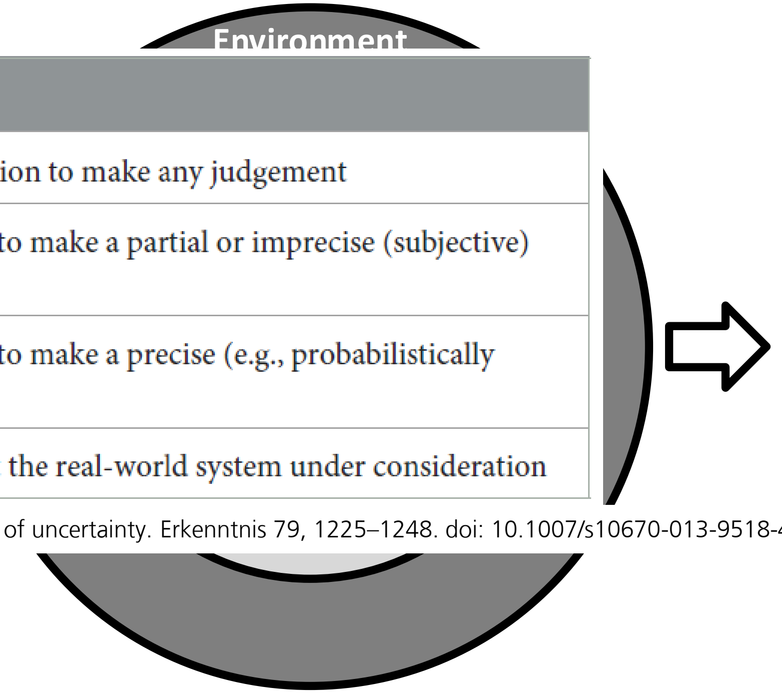
Any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system

Level	Definition
Level 4	Knowledge K of structural relationships of the system under consideration can not be assumed. It may, however, be possible to rank K subjectively such that higher uncertainty is entailed in a lower ranking of K.
Level 3	Uncertainty refers to the completeness of the evidence on which the judgement of probability is reached. *Weight* is a measure of completeness of relevant evidence. On this level, subjective probabilities or evidence theory may be useful. It can thus be seen as referring to the *validity* of available evidence.
Level 2	Uncertainty is represented as a matter of belief and is inversely proportional to the probability measure, i.e., it is greater, the lower the probability measure becomes. It can thus be measured by $1 - p$ where p is the degree of belief in the argument a conditional on evidence h. An important measure here are statistical confidence intervals. Levels 1 and 2 can be viewed as referring to the *integrity* of available evidence.
Level 1	Uncertainty is inherent in reality and can be captured in a stochastic term ϵ. The degree of uncertainty is then measured by the variance of ϵ, i.e., $\sigma(\epsilon)$.

Dow, S. C. (2012). Uncertainty about Uncertainty. London: Palgrave Macmillan UK.

Severity	Definition
Ignorance	Not enough information to make any judgement
Severe	Enough information to make a partial or imprecise (subjective) judgement
Mild	Enough information to make a precise (e.g., probabilistically correct) judgement
Certainty	Full knowledge about the real-world system under consideration

Bradley, R., and Drechsler, M. (2014). Types of uncertainty. Erkenntnis 79, 1225–1248. doi: 10.1007/s10670-013-9518-4



Assurance Uncertainty:
Lack of confidence in assurance arguments

Manifestations of uncertainty

Interpretation of: Lovell, B. E. (1995). A taxonomy of types of uncertainty

Overview of the presentation

1. Introduction and motivation
2. Defining uncertainty
3. Uncertainty and machine learning from a safety perspective
4. Constructing a safety assurance argument for machine learning
5. Analysing uncertainties in the assurance argument
6. Continuous assurance
7. Outlook and research perspectives

Uncertainty and machine learning

Specification insufficiencies

Arguing the safety of machine learning functions requires:

A detailed and measurable specification of the safety requirements:

- E.g.: Each pedestrian within the **critical range** is correctly detected within any sequence of **N images** with a **true positive rate**, **vertical** and **horizontal deviation** from ground truth sufficient to avoid collisions.
- Which KPIs/Metrics can be used to measure the conformance to the requirements?
- How to derive threshold values (validation targets) for these metrics?

A detailed understanding of the operational design domain and system context:

- E.g.: Distribution various types of pedestrians, definition of critical scenarios, capabilities of sources (e.g. camera) and consumers (e.g. planning algorithms) within the system context.

Beware of the specification paradox!

- If we use of ML to learn “unspecifiable” behaviour via a set of representative training data, how do we define under which set of conditions the function is safe? and what are the consequences of using data from the same distribution to do so?

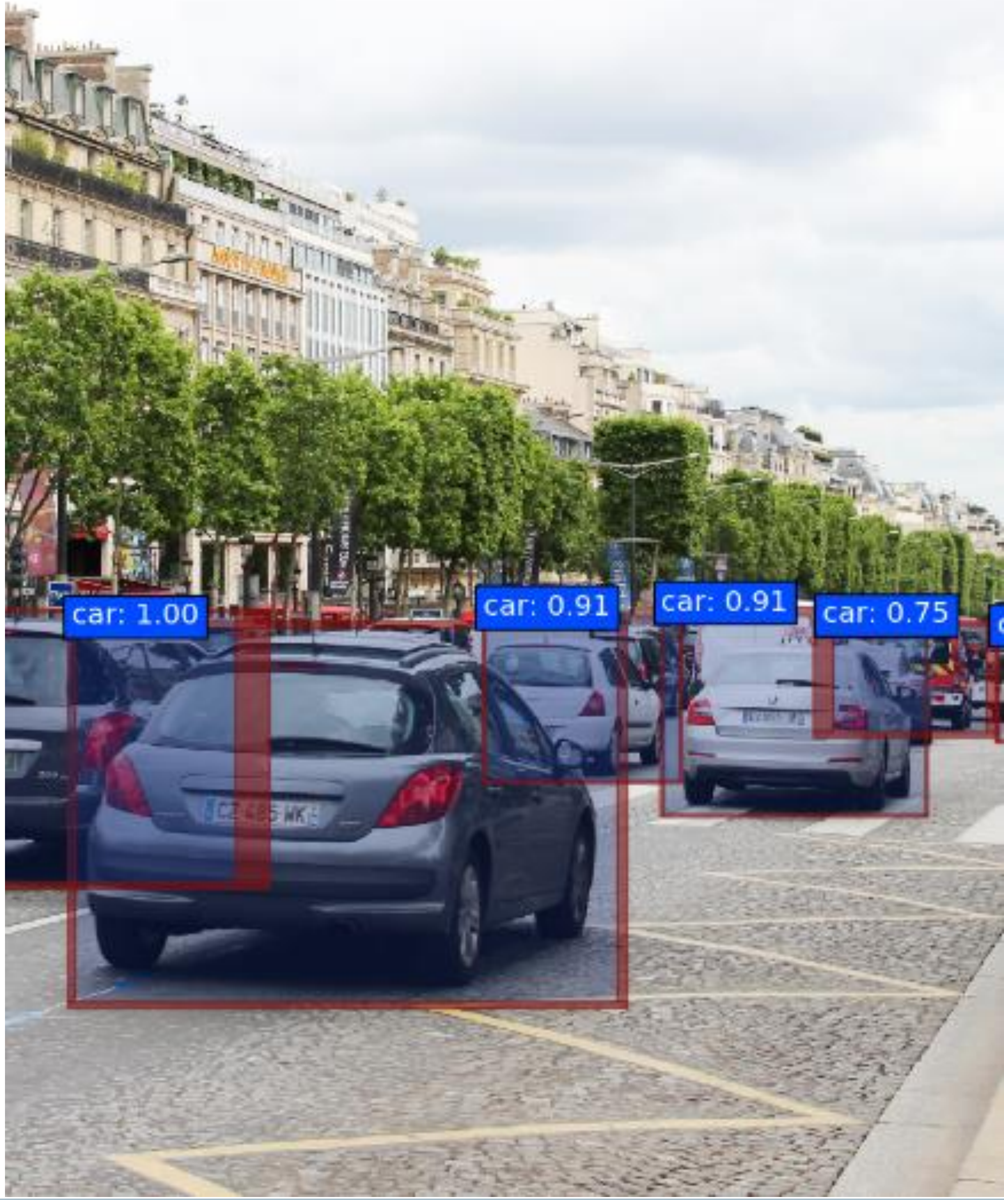
Uncertainty and machine learning
Performance insufficiencies

Machine Learning can be seen as a class of heuristic algorithms:

- **Heuristic:** technique for solving a problem more quickly when classic methods are too slow for finding an approximate solution, or when classic methods fail to find any exact solution. This is achieved by **trading optimality, completeness, accuracy, or precision** for speed.

Gaps between theoretically optimal function and the trained model

- Robustness, generalization, Bias: outputs sensitive to small changes in the inputs, semantic deficiencies in training data, …
- Prediction uncertainty: Confidence scores not necessarily indication of probability of correctness
- Related to the concepts of **task complexity/learnability**, **sample complexity** (number of samples required for a problem to be efficiently learnable) and **model expressiveness** (inherent capacity of the model to express functions of a given complexity)



Uncertainty and machine learning

Definition of the safety assurance problem

We would like to demonstrate that for all inputs ***i of I***, the model ***M*** fulfils its safety guarantees ***G,*** under the assumptions ***A***

$$\forall i \in I.A(i) \Rightarrow G(i, M(i))$$

Absolute perfection is neither achievable nor required to achieve a tolerable level of residual risk according to an acceptance criteria (***AC***), therefore we need to achieve a ***probability of success*** for a ***given distribution of inputs*** in the operational design domain ***ODD***

$$\frac{\sum_{i \in I, A(i) \wedge G(i, M(i))} \mathbb{P}_{ODD}(i)}{\sum_{i \in I, A(i)} \mathbb{P}_{ODD}(i)} \geq AC$$

But, we cannot directly demonstrate the Guarantees ***G***, for all inputs. Instead, we can evaluate measurable properties ***P*** of ***M*** (e.g. precision, recall, robustness, calibrated error rate) for a finite number of samples ***j*** of ***I*** (e.g. our test dataset)

Definition of the safety assurance problem:
How can we argue that a sufficiently small residual risk has been achieved, despite potential insufficiencies in the specification and performance of the ML function, based on an **appropriate selection** of measurable properties ***P*** and input samples ***j***?

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

Which combination of Properties ***P*** best represent ***G***?

Which samples ***j*** are representative of the input domain?

Fraunhofer IKS

Overview of the presentation

1. Introduction and motivation
2. Defining uncertainty
3. Uncertainty and machine learning from a safety perspective
4. Constructing a safety assurance argument for machine learning
5. Analysing uncertainties in the assurance argument
6. Continuous assurance
7. Outlook and research perspectives

Safety assurance arguments

What is assurance?

assurance

grounds for justified confidence that a *claim* has been or will be achieved

assurance case

reasoned, auditable artefact created that supports the contention that its top-level *claim* (or set of claims) is satisfied, including systematic *argumentation* and its underlying *evidence* and explicit *assumptions* that support the claim(s)

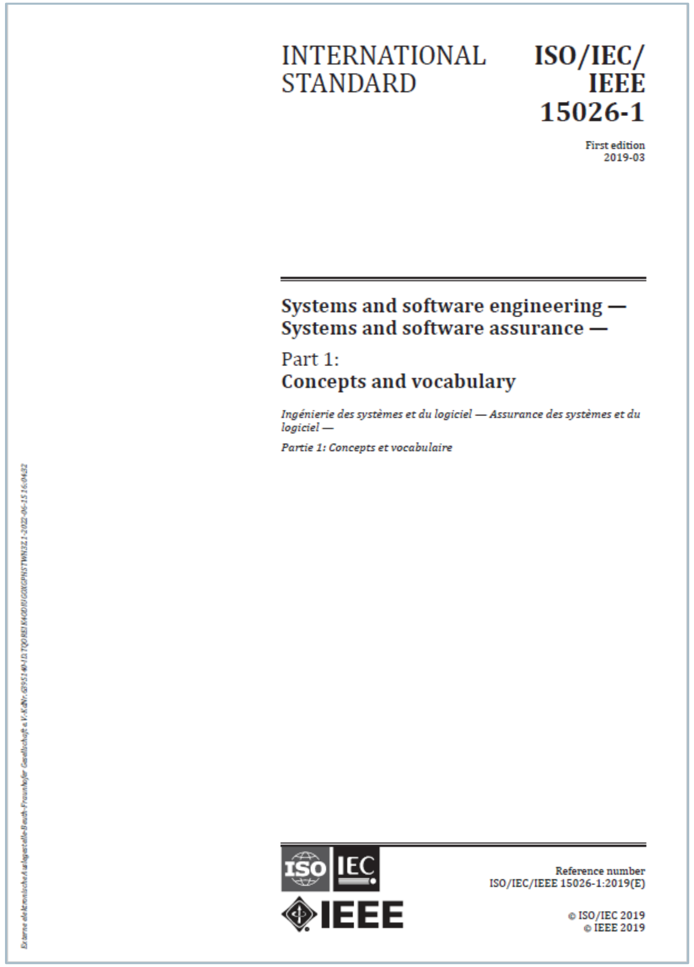
Note 1 to entry: An assurance case contains the following and their relationships:

— one or more claims about properties;

— arguments that logically link the evidence and any assumptions to the claim(s);

— a body of evidence and possibly assumptions supporting these arguments for the claim(s); and

— justification of the choice of top-level claim and the method of reasoning.


INTERNATIONAL STANDARD

ISO/IEC/ IEEE 15026-1

First edition
2019-03

Systems and software engineering — Systems and software assurance —

Part 1:
Concepts and vocabulary

Ingénierie des systèmes et du logiciel — Assurance des systèmes et du logiciel —

Partie 1: Concepts et vocabulaire

Reference number
ISO/IEC/IEEE 15026-1:2019(E)

© ISO/IEC 2019
© IEEE 2019

Safety assurance arguments

Modelling assurance arguments / safety cases

Goal Structuring Notation (GSN)[1]

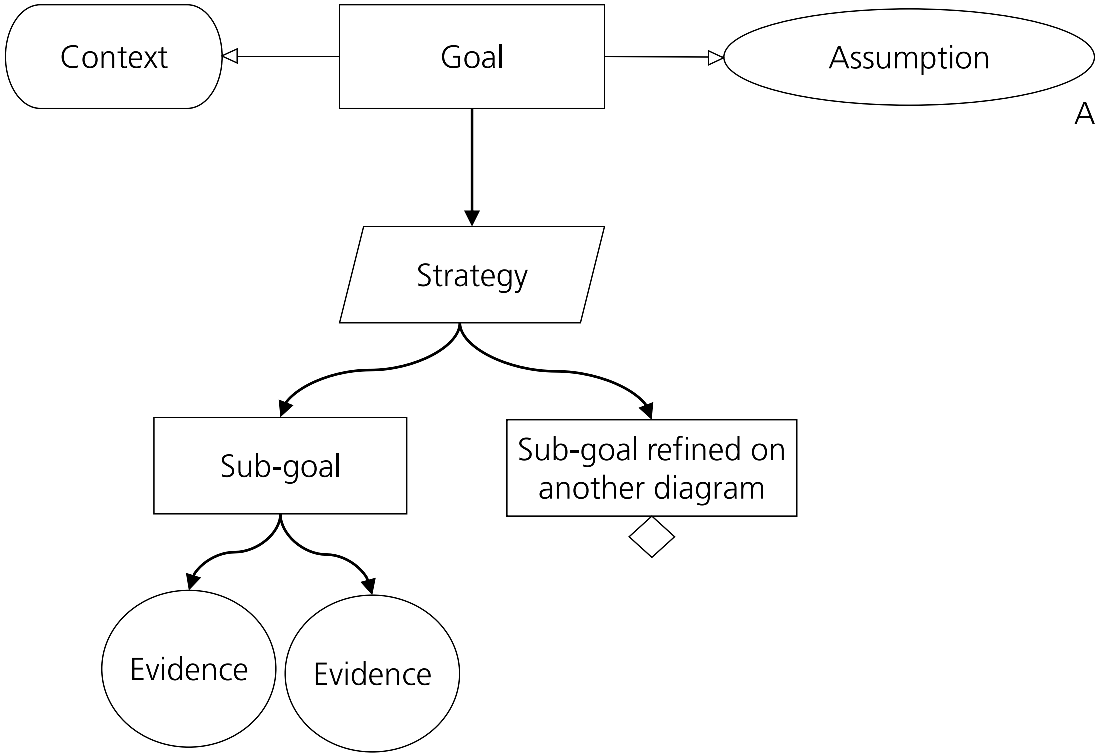
Graphical notation that represents the elements of an assurance argument and the relationships between them

Shows how **goals** (claims) can be broken into **sub-goals** until they can be supported by direct reference to **evidence**

Documents argumentation **strategies** as well as **context** information, including **assumptions** and **justifications**

Can be structured hierarchically and modularly, assurance claim points used to indicate where additional argumentation is required to increase confidence in the argument

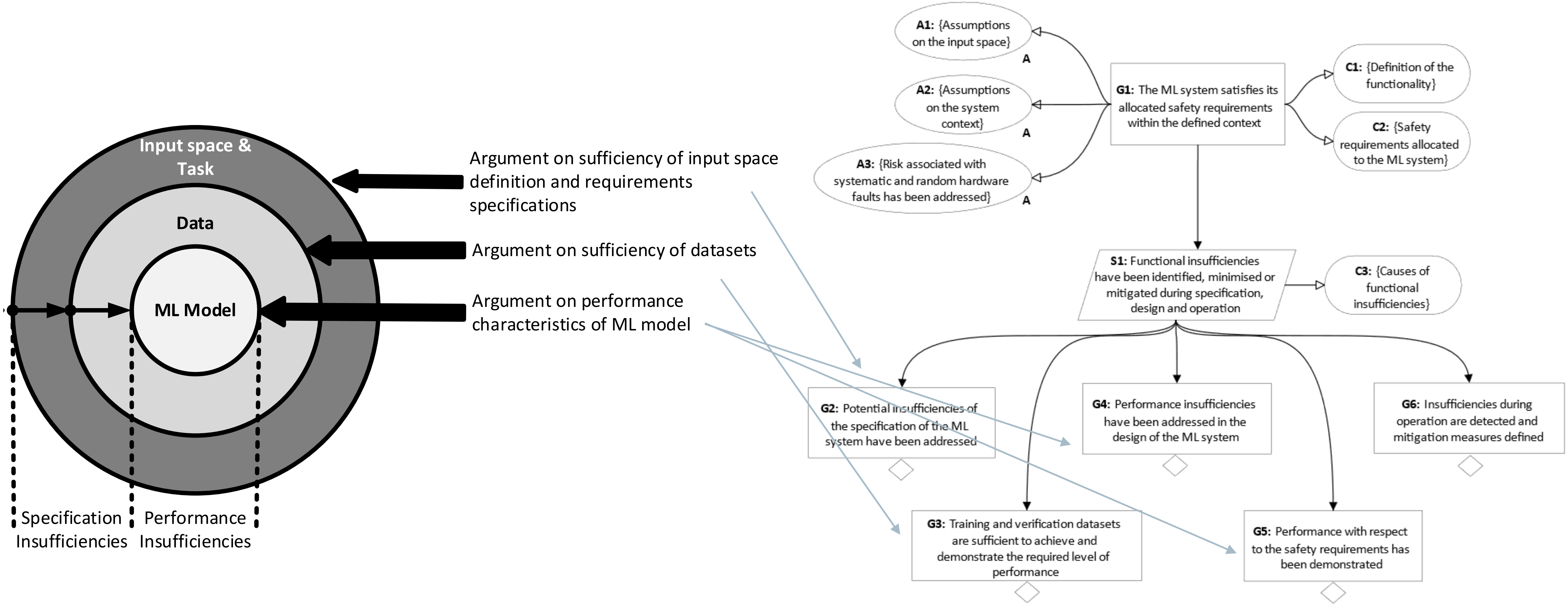
Defined uses the Structured Assurance Case Metamodel (SACM)[2]



[1] https://scsc.uk/gsn
[2] https://www.omg.org/spec/SACM

Safety assurance arguments

Assurance arguments for machine learning



Safety assurance arguments

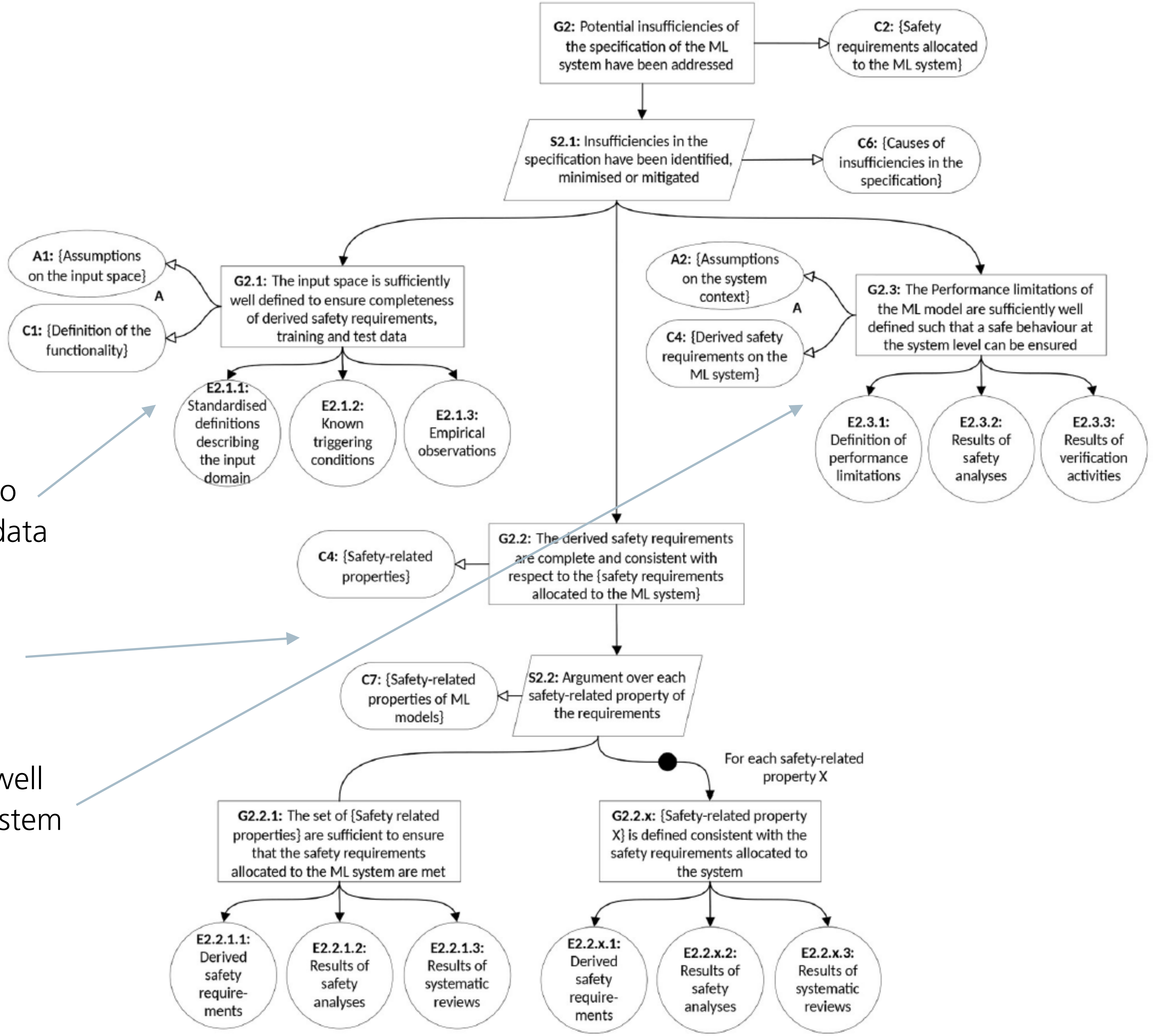
Assurance arguments for machine learning

Example elaboration of the assurance argument regarding potential insufficiencies in the specification

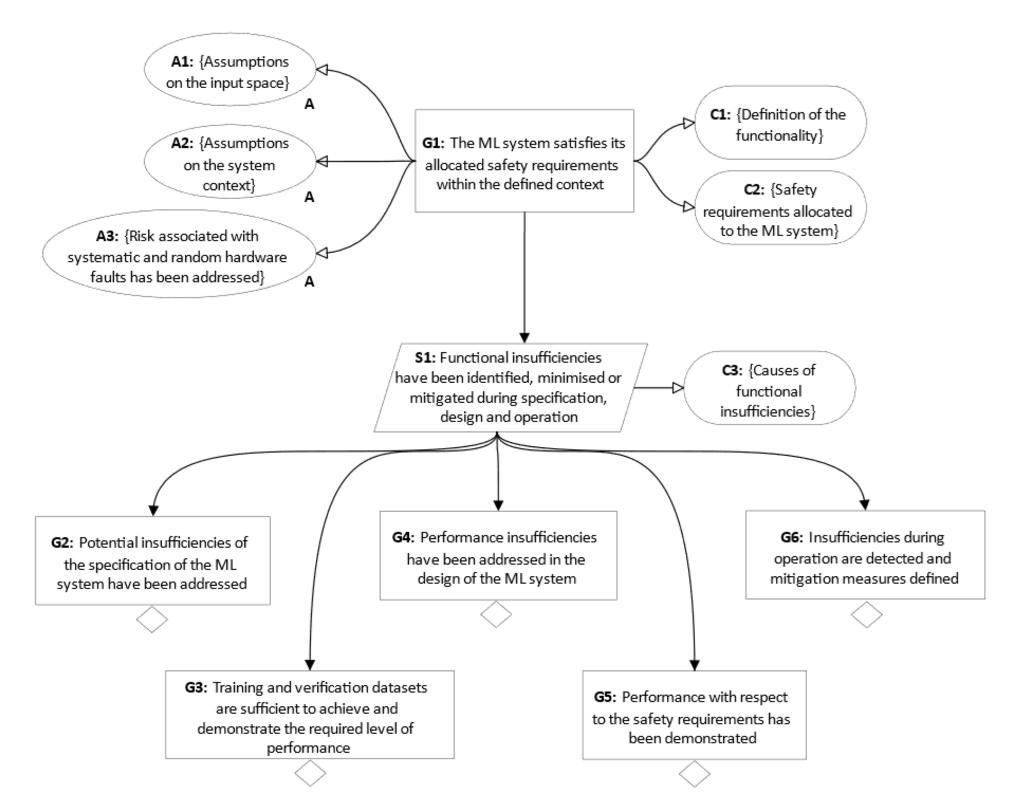
The input space (ODD) is sufficiently well understood and defined to ensure completeness of the safety requirements, training and test data

The set of derived requirements and properties used to define validation targets is sufficient to ensure that the higher-level safety requirements allocation to the ML function are fulfilled.

Potential performance limitations of the ML model are sufficiently well defined, such that residual errors can be compensated for at the system level



Even if we take a structured approach to formulating a safety assurance argument for ML – can we really trust the argument to reflect the actual residual risk of the system?



Overview of the presentation

1. Introduction and motivation
2. Defining uncertainty
3. Uncertainty and machine learning from a safety perspective
4. Constructing a safety assurance argument for machine learning
5. Analysing uncertainties in the assurance argument
6. Continuous assurance
7. Outlook and research perspectives

Fraunhofer IKS

Confidence in assurance arguments

Types of assurance uncertainty

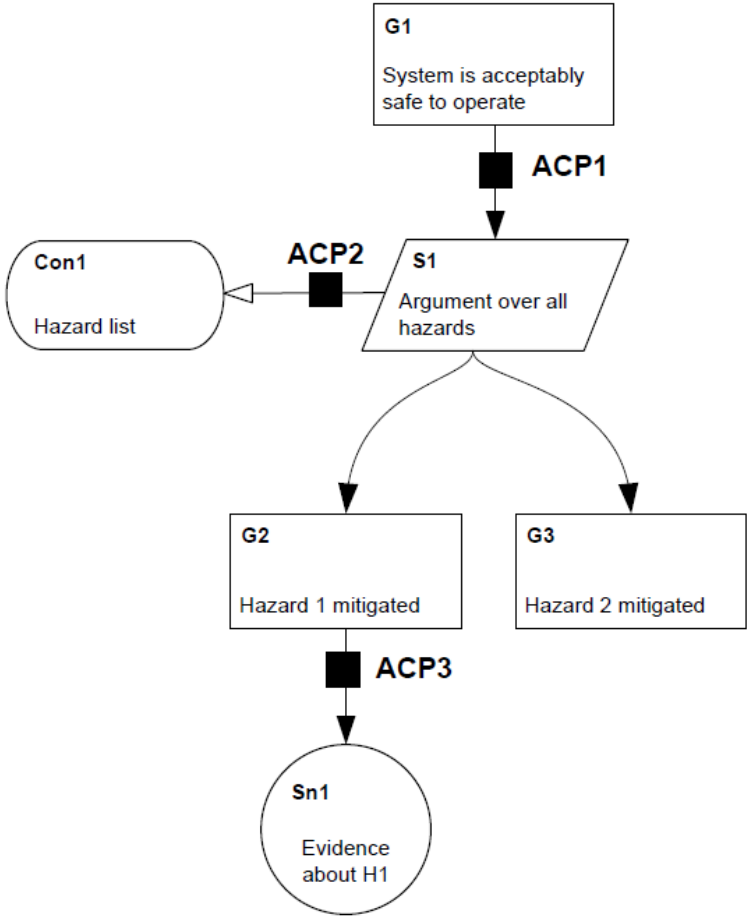
Assurance Claim Points (ACPs)

Indicate **assertions** in the Assurance argument whose truth must be justified in order for the argument to be compelling (i.e. believed).

Confidence arguments can be used to provide the justification for the assertions

Allows for a separation of concerns when developing the assurance argument:

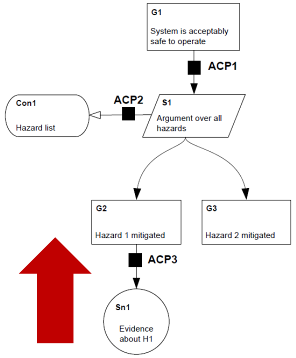
- ***Arguing properties of the product vs. arguing properties of the argument!***



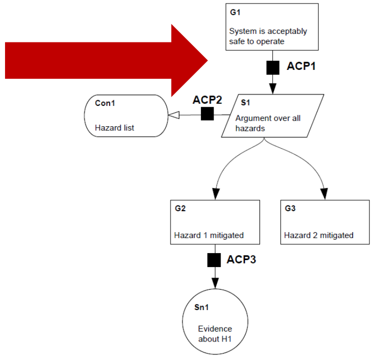
Hawkins, Richard, Tim Kelly, John Knight, and Patrick Graydon. "A new approach to creating clear safety arguments." In *Advances in systems safety*, pp. 3-23. Springer, London, 2011.

Confidence in assurance arguments

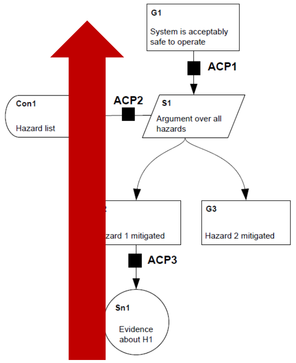
Assurance Claim Points



Asserted evidence: The evidence that is put forward is sufficient to support the claim and is trustworthy.



Asserted context: Context (e.g. assumptions) is appropriate for the argument elements (e.g. claims) to which it applies.



Asserted Inference: Probable truth of the premises (sub-claims) is sufficient to establish the probable truth of the conclusion (Claim).

Hawkins, Richard, Tim Kelly, John Knight, and Patrick Graydon. "A new approach to creating clear safety arguments." In *Advances in systems safety*, pp. 3-23. Springer, London, 2011.

Uncertainty in the assurance arguments for ML

Asserted context

Asserted context – Assumptions on the input space

Are all assumptions on the input space to the ML function valid?

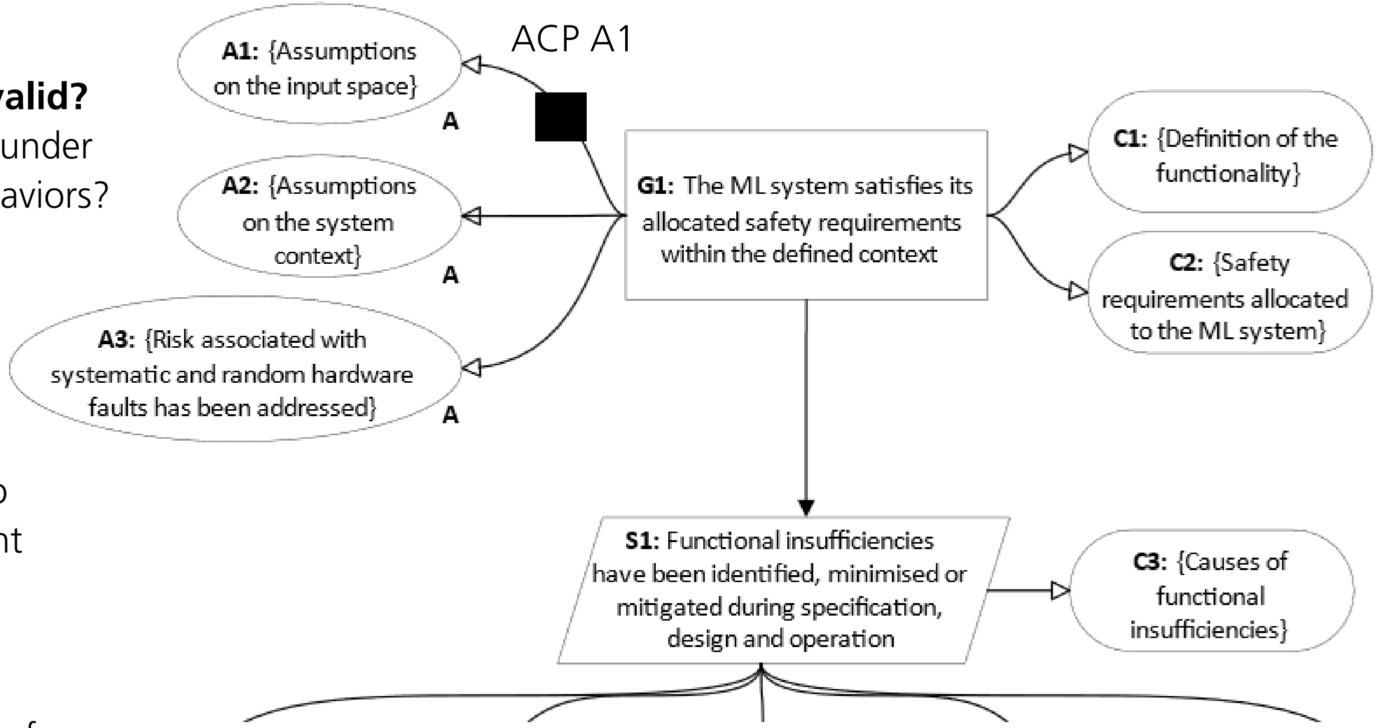
- E.g. is there consensus as to what constitutes a pedestrian and under which conditions pedestrians could appear and with which behaviors?
- Implicit assumption on the input space would undermine the confidence in the argument that the safety requirements have been adequately defined

Severity of uncertainty:

- Only qualitative definition of the input space possible leading to **severe uncertainty** and possibly **ignorance** (level 4) of relevant characteristics of the pedestrians or the environment

Improvement measures:

- Simplification of requirements to detect all obstacles regardless of human or non-human, more restrictive assumptions on the operational design domain, …



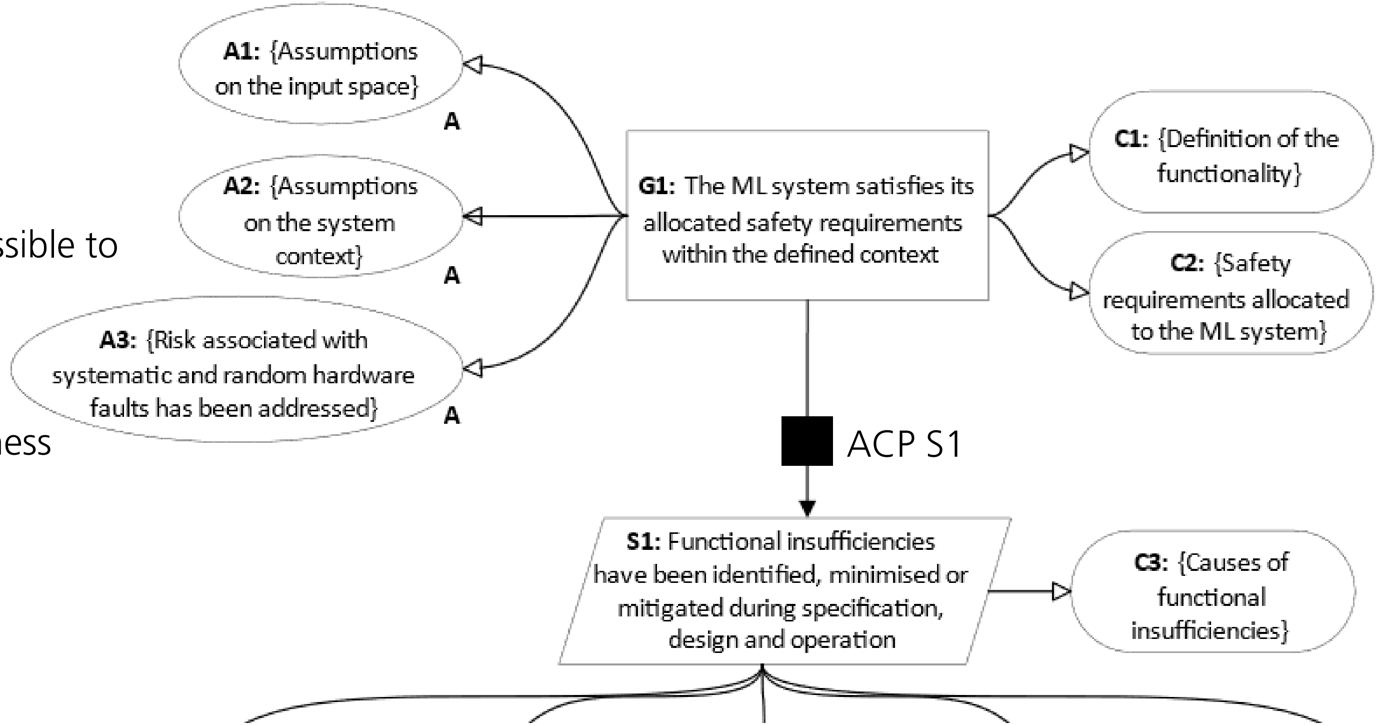
Uncertainty in the assurance arguments for ML

Asserted inference

Asserted inference – Completeness of the argument

Have all possible causes of functional insufficiencies been addressed?

- Due to the inherent complexity of the environment and system (including the black-box nature of ML itself) it might not be possible to directly identify causes of functional insufficiencies
  - partial observability of failure causes due to entanglement of causal factors of insufficiencies
  - E.g. Relevance of "Out-of-distribution" events and effectiveness of out-of-distribution detection measures



Severity of uncertainty:

- **Severe observational uncertainty** regarding the causes of insufficiencies

Improvement measures:

- Systematic safety analysis supported by targeted experiments

Uncertainty in the assurance arguments for ML

Asserted evidence

Asserted evidence – integrity and validity of verification results

To what extent does a particular verification evidence imply that a property of the ML model has been achieved?

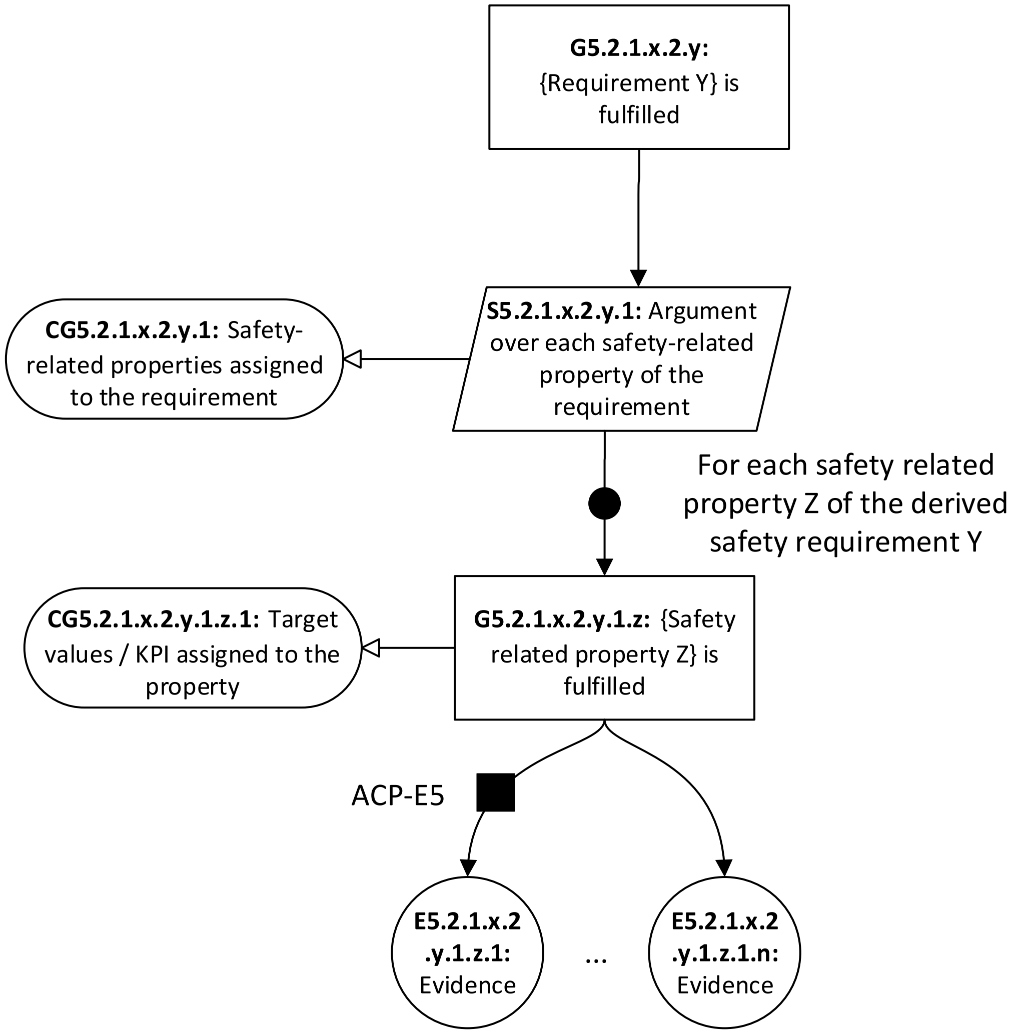
- Are the verification results representative of the actual performance in the field?
- Have representative samples been used, and has a sufficient coverage of the input space been achieved?
- Have the results of the verification activities been correctly interpreted?

Severity of uncertainty:

- E.g. what level of statistical confidence has been achieved with the evidence (level 2)?
- Are the assumptions used to extrapolate the results of the tests valid (level 3 and 4)?

Improvement measures:

- Statistical analysis of test results, diverse verification methods (e.g. ODD coverage *and* search-based testing for corner cases), additional arguments for confidence in the evidence



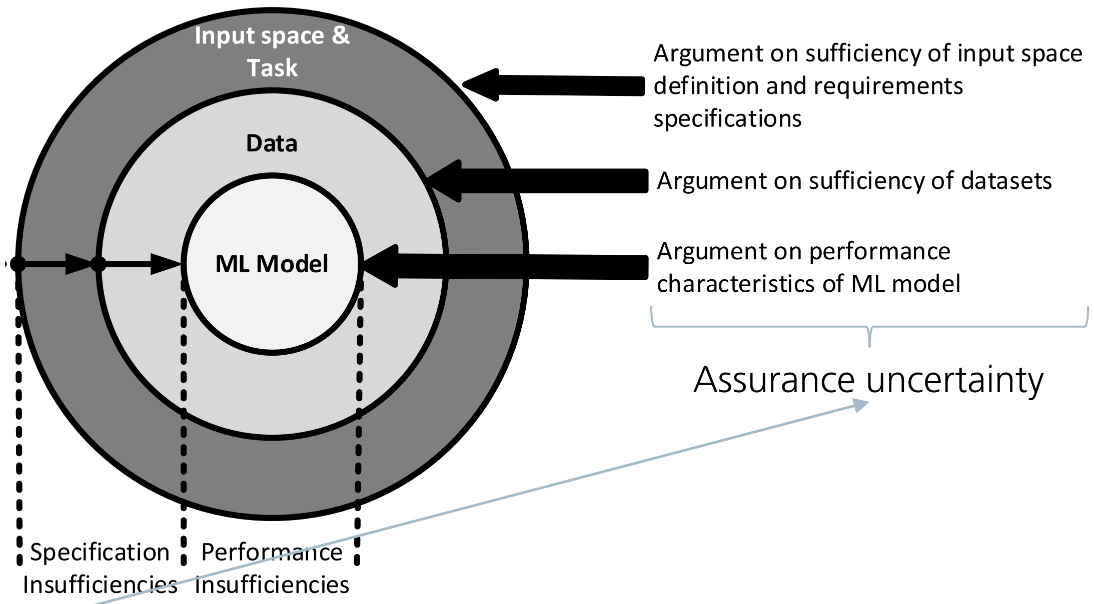
Overview of the presentation

1. Introduction and motivation
2. Defining uncertainty
3. Uncertainty and machine learning from a safety perspective
4. Constructing a safety assurance argument for machine learning
5. Analysing uncertainties in the assurance argument
6. Continuous assurance
7. Outlook and research perspectives

Residual uncertainty in the assurance argument

Definition of the safety assurance problem:
How can we argue that a sufficiently small residual risk has been achieved, despite potential insufficiencies in the specification and performance of the ML function, based on an **appropriate selection** of measurable properties ***P*** and input samples ***j***?

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

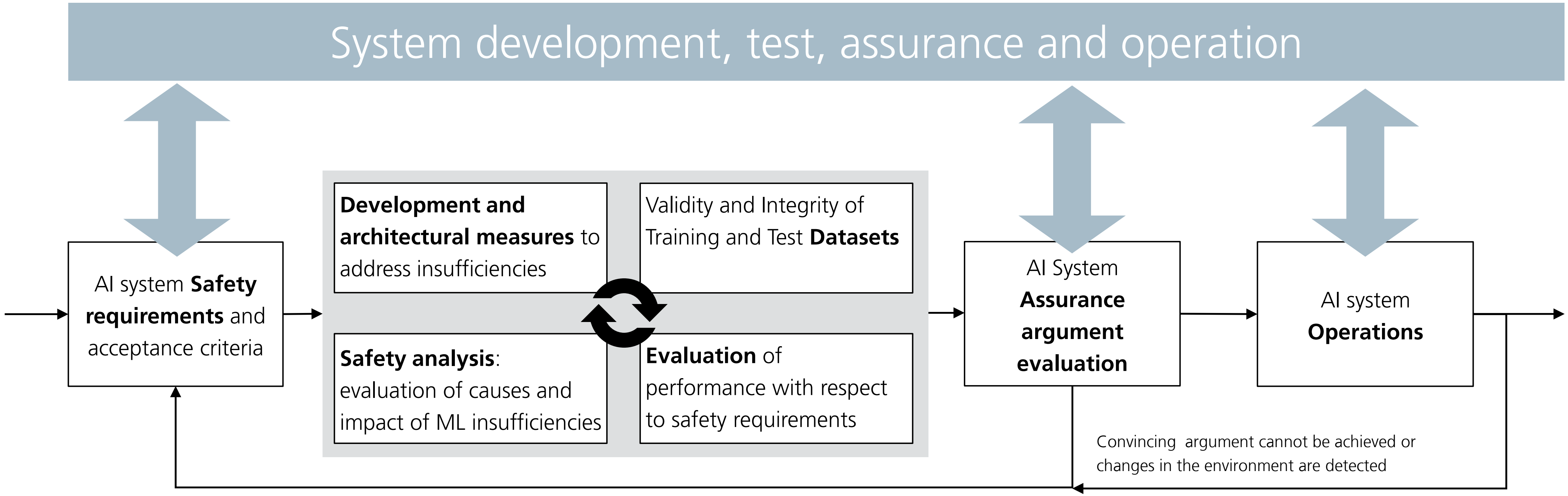


For any non-trivial system there will inevitably be a gap between our estimated and the actual achieved level of risk

This needs to be compensated for by either conservative methods of assurance and/or mitigations at the system level

Continuously reducing uncertainties

Safety Lifecycle for AI/ML-based functions



Uncertainties in the specification, models and the assurance case must be iteratively reduced over time

The level of environment, task and system complexity must be increased in line with increasing confidence

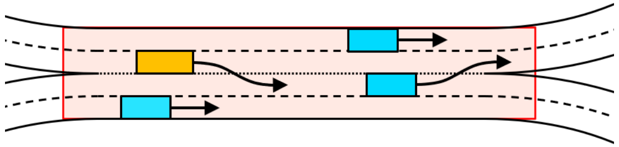
Overview of the presentation

Discussion: will ML ever be „Safe enough“?

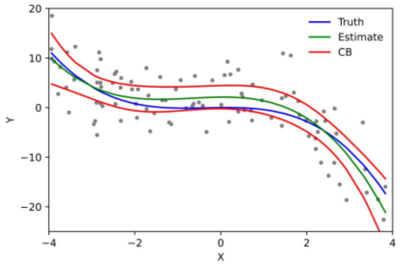
Yes, but for which levels of complexity and definition of „safe enough“?

Uncertainty in the assurance argument

- E.g.: DNNs for Pedestrian prediction



- E.g.: Support Vector Machines for trajectory prediction



- E.g.: Linear Regression for „virtual sensors“ in control algorithms

Complexity of environment, task, system and models

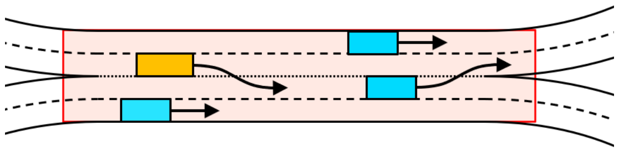
Discussion: will ML ever be „Safe enough“?

Yes, but for which levels of complexity and definition of „safe enough“?

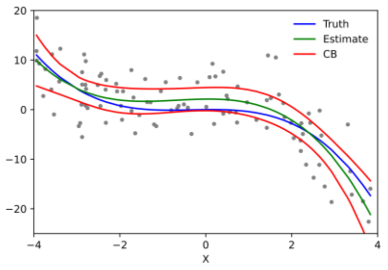
Uncertainty in the assurance argument

Development of state-of-the-art in safety assurance methods, Application of ISO PAS 8800

- E.g.: DNNs for Pedestrian prediction



- E.g.: Support Vector Machines for trajectory prediction



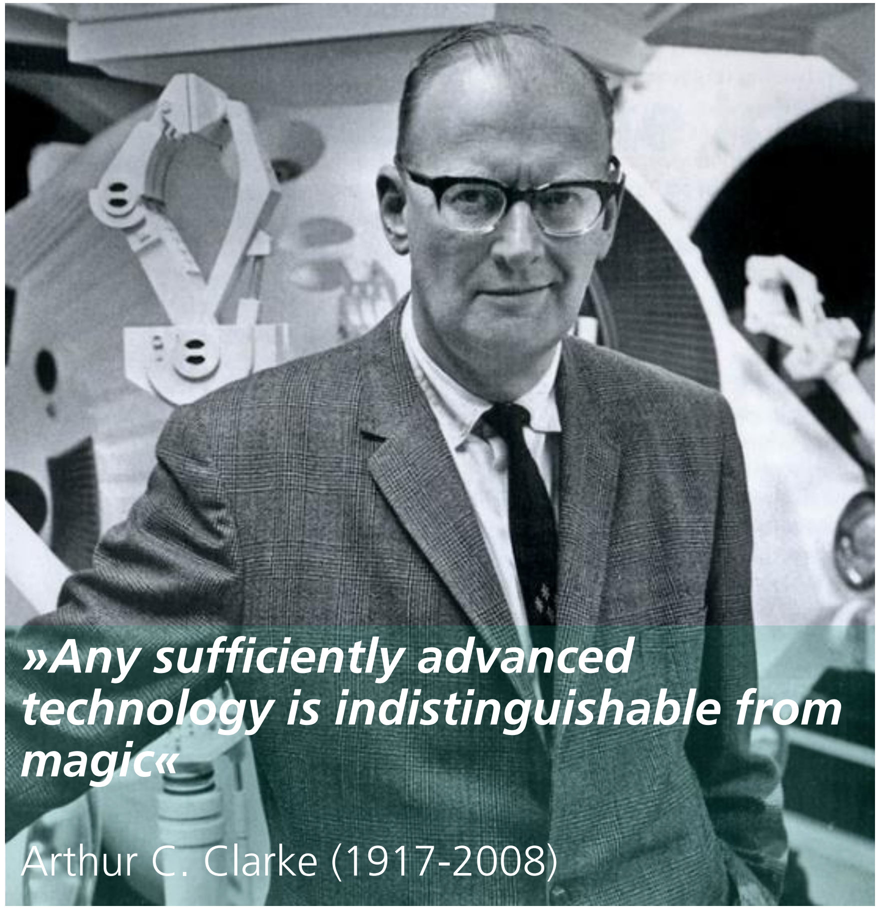
- E.g.: Linear Regression for „virtual sensors“ in control algorithms

Complexity of environment, task, system and models

Safety assurance of AI/ML

Some ongoing research questions

- Bridging the gap between societal and ethical expectations and technical acceptance criteria
- Definition of risk acceptance criteria for complex highly-automated systems, including target values for common ML metrics
- Addressing assurance uncertainty: Role of quantitative and qualitative evidence in assuring the safety of cognitive cyber-physical systems
- Engineering of "Safe" and "Assurable" ML approaches
- Continuous, automated safety assurance of AI/ML
- Uncertainty propagation analysis during design and run-time uncertainty quantification
- Safety assurance of self-adaptive systems



Addressing uncertainty in assurance of ML

Summary

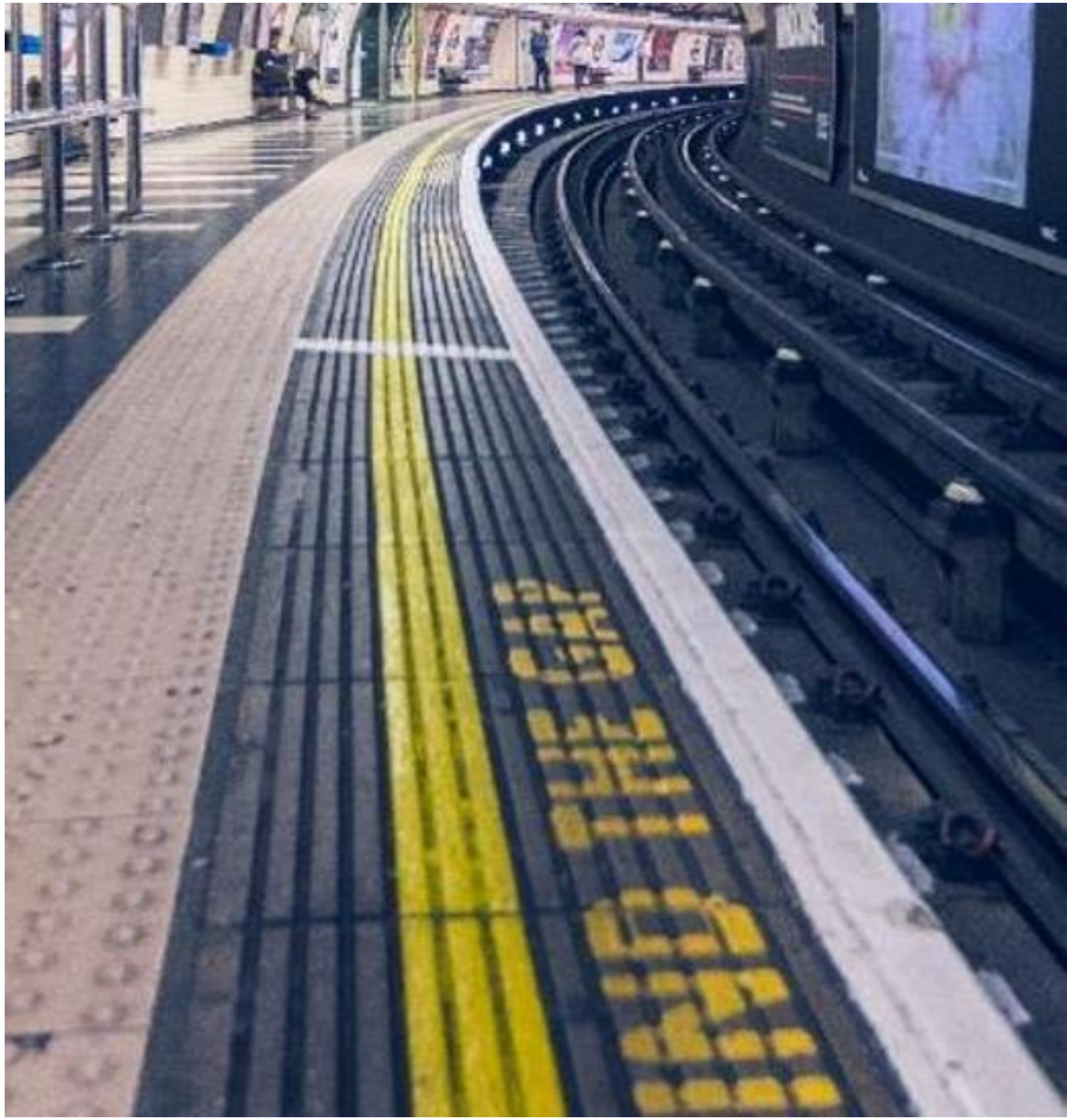
There are no straightforward answers to whether ML is safe or not, as this depends on:

- Our ability to precisely express the safety conditions of the task
- The nature of the data used to train and test the system
- Underlying properties of the technologies and algorithms used

A systematic approach is required to reason about safety of ML

But, we also need to understand, and compensate for, the limits of our safety argumentation

The use of assurance arguments and assurance claim points, along with an understanding of the causes and manifestations of uncertainty provide a promising way forward



Contact

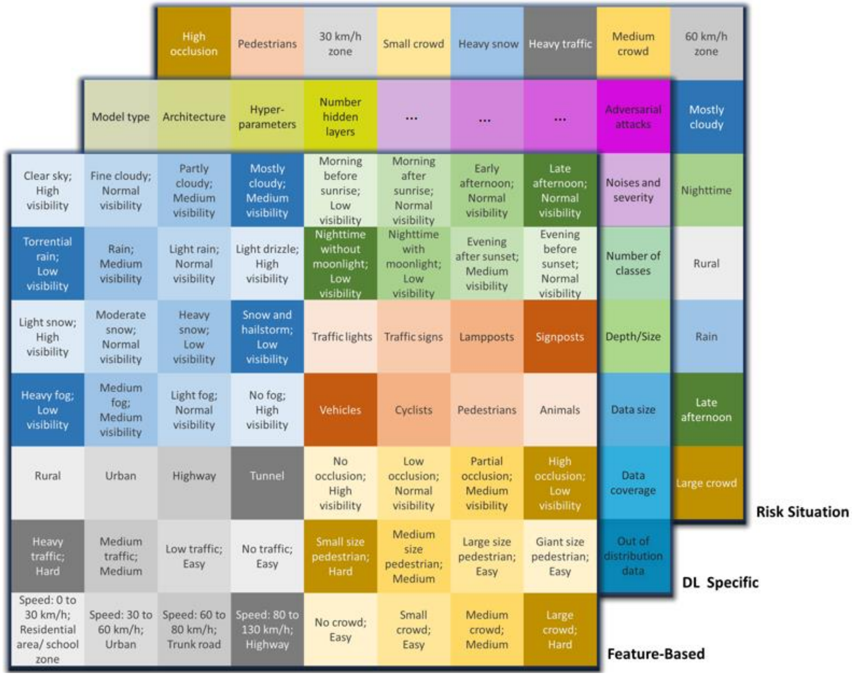
Prof. Simon Burton
Research Division Director, Safety Assurance
Tel. +49 89 547088-341
simon.burton@iks.fraunhofer.de

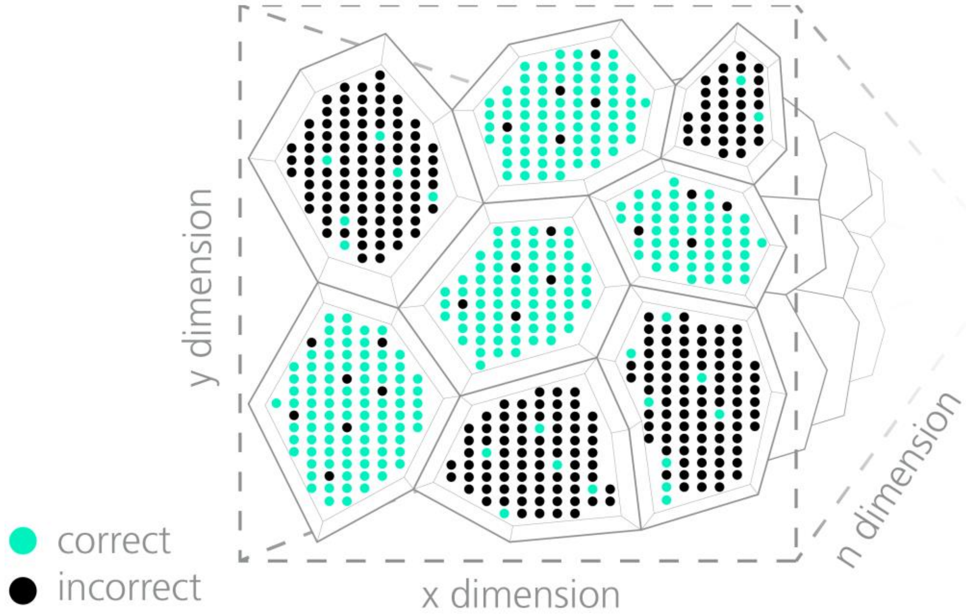
Fraunhofer IKS
Hansastrasse 32
80686 München
www.iks.fraunhofer.de

Other ongoing research

Micro Operational Design Domains (µODDs)

Organising systematic factors influencing the perception system





Concept of µODDs allows to split ODD into several µODDs (e.g., for benefiting from different levels of risk in different situations*)

Use of µODDs to describe an operational condition in which the occurrence of a ML error can be treated as aleatoric uncertainty

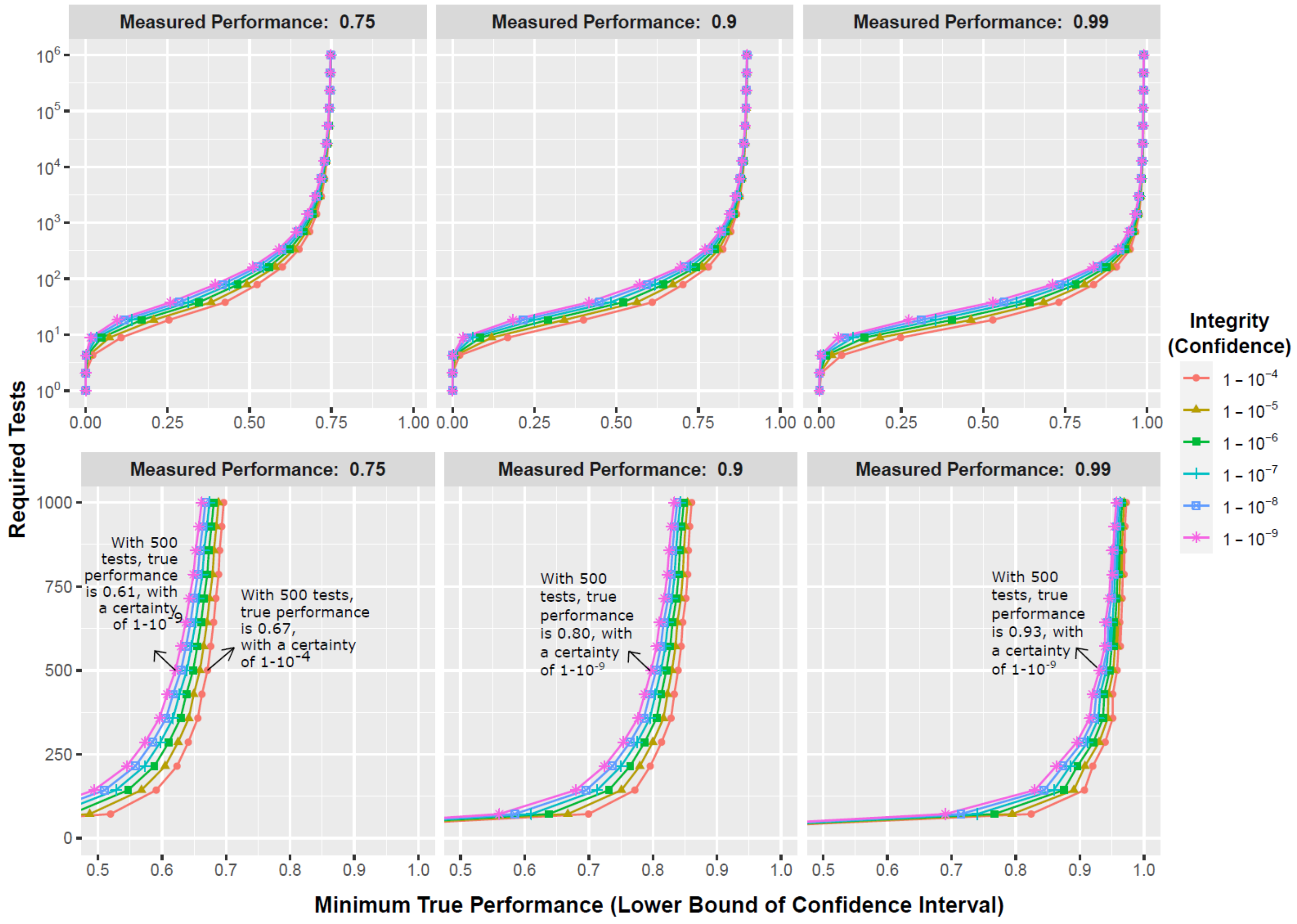
*P. Koopman, B. Osyk, and J. Weast, "Autonomous vehicles meet the physical world: RSS, variability, uncertainty, and proving safety," in Extended Preprint of Int. Conf. on Computer Safety, Reliability, and Security, 2019, pp. 245–253. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1911/1911.01207.pdf

Source: P. Schleiss, Y. Hagiwara, I. Kurzidem, F. Carella: „Towards the Quantitative Verification of Deep Learning for Safe Perception", in Proc. of 12th IEEE International Workshop on Software Certification at 33rd IEEE International Symposium on Software Reliability Engineering (ISSRE), 2022.

Quantitative verification of AI

Relationship between test sample size, measured performance and confidence

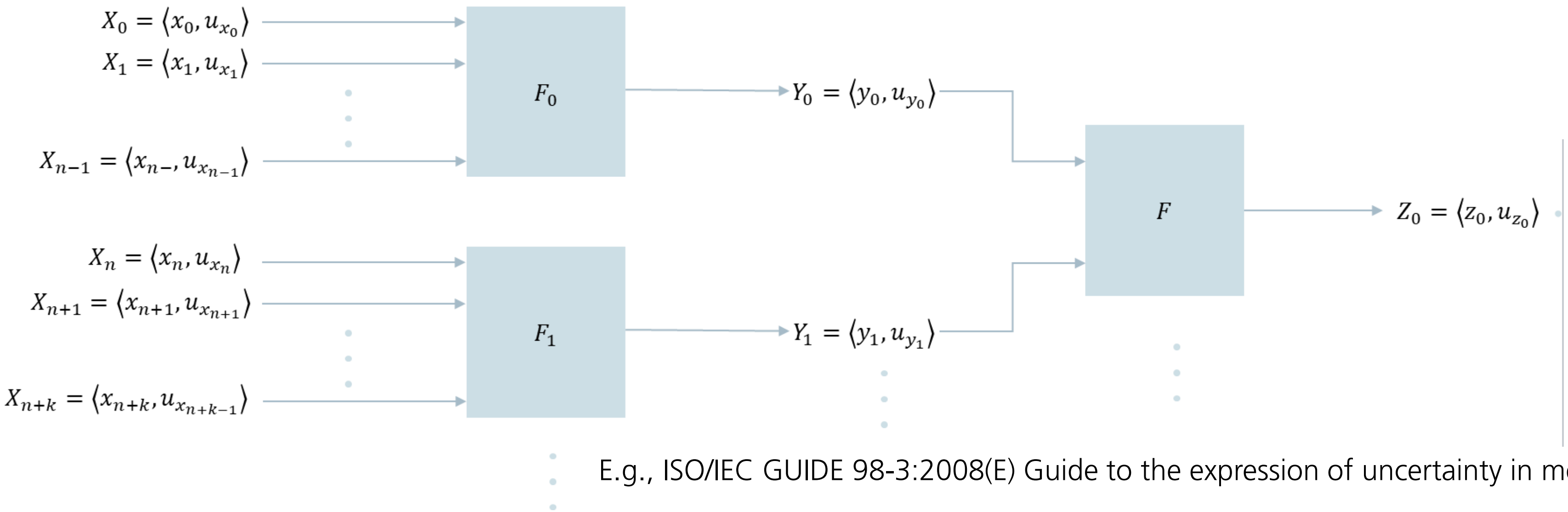
- **Confidence (probability of not making a statistical error) generally grows with increased sample size**
- **Distance between measured performance and required minimal performance in the field also influences confidence**
- **Example**
  - Binomial tests are used for classification problem
  - When measuring a performance of 75% after 500 tests the minimal performance in the field will be at least 61% with a confidence of 99.9999999% or even 67% with a reduced confidence of 99.99% (bottom left)
- **Concept may not scale when facing a multitude of μODDs**
- **What about other metrics? E.g. robustness, heat-maps, etc.**

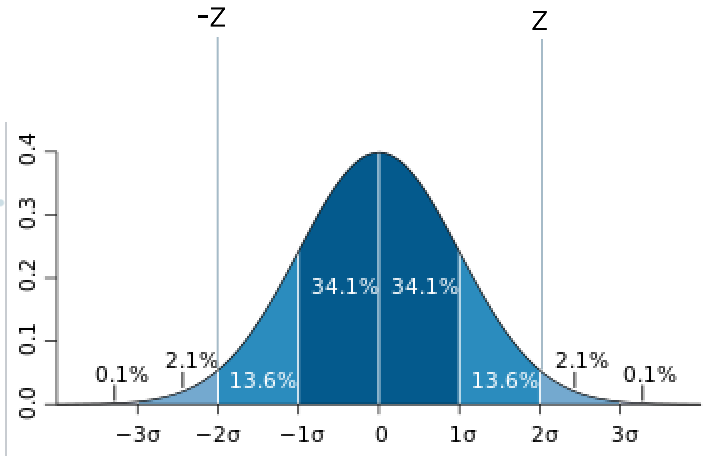


  Public information

Measuring Uncertainty @ Runtime

Increasing utility by consider current level of risk instead of worst case considerations





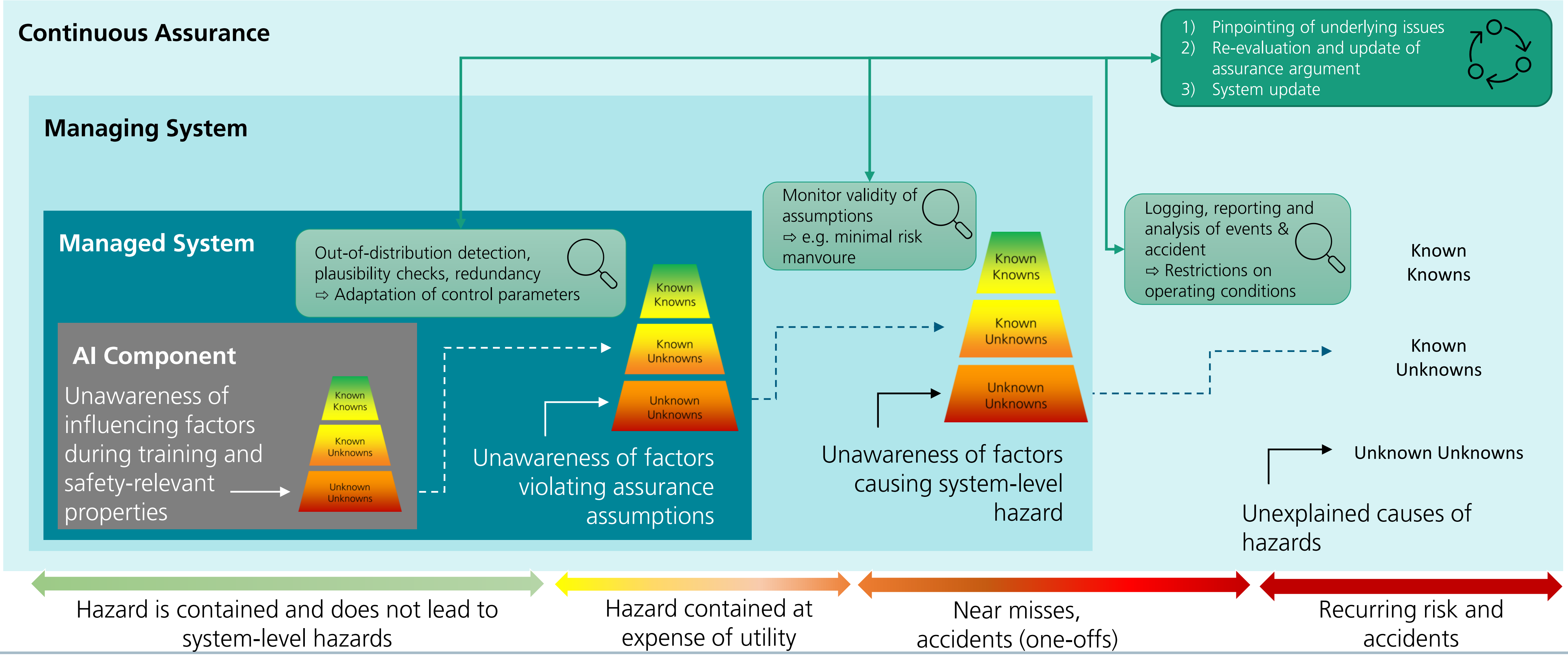
E.g., ISO/IEC GUIDE 98-3:2008(E) Guide to the expression of uncertainty in measurement

Uncertainty quantification and propagation:

- First research for quantifying ML- uncertainty at runtime beyond soft-max (e.g., deep ensembles & out-of-distribution detection)
- 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**
- Statistical soundness critical for correct uncertainty propagation and estimation at system level

Limits of Quantifying Uncertainty

Layers of uncertainty: How to benefit from continuous assurance?



Fraunhofer IKS