Fraunhofer ESK

Dominique Seydel

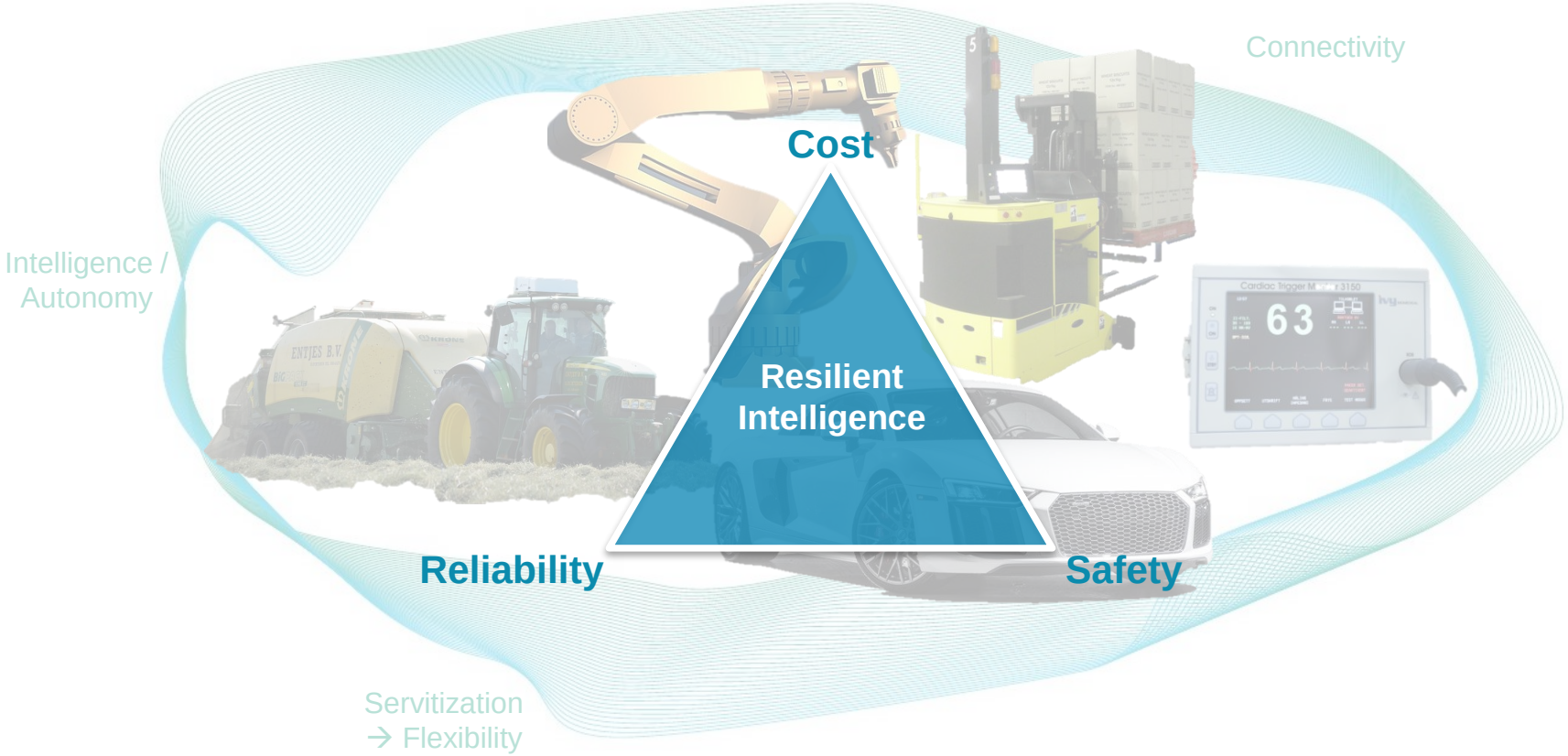
Research Engineer
dominique.seydel@esk.fraunhofer.de

SAFE INTELLIGENCE

USE OF AI-BASED SOLUTIONS IN SAFETY-CRITICAL APPLICATIONS

KOGNITIVE SYSTEME | KÜNSTLICHE INTELLIGENZ & MACHINE LEARNING | AUTONOME SYSTEME | AUTONOMES FAHREN | INDUSTRIE 4.0 | IOT

EVOLUTION OF VEHICLES AND MACHINES TOWARDS AUTONOMOUS SYSTEMS



HOW TO VALIDATE SAFETY FOR AI-BASED SYSTEMS?

WHAT ARE THE CURRENT WEAKNESSES OF AI?

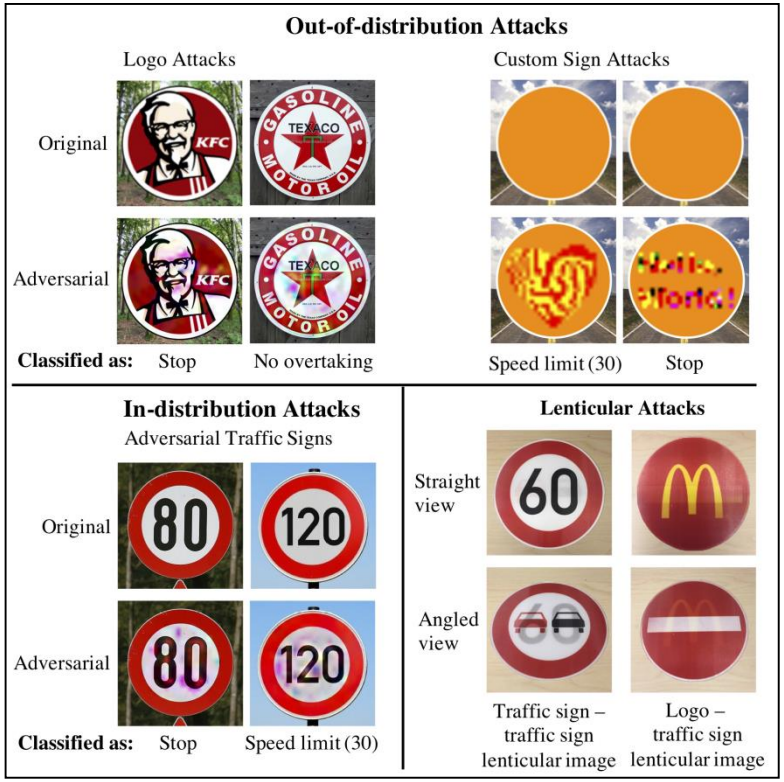
Limited capabilities of today's Neural Networks

- **No generalization** (no separation of essential aspects of an object from the context)
- **No extrapolation** (no abstract concept transferrable to other subject areas)
- **No global model** (only statistical no causal correlations)
- **Results not comprehensible** (significance of variables not legible)
- Learned **behavior is non-deterministic** (small modifications of input lead to completely different output classifications)

WHAT ARE THE CURRENT WEAKNESSES OF AI?

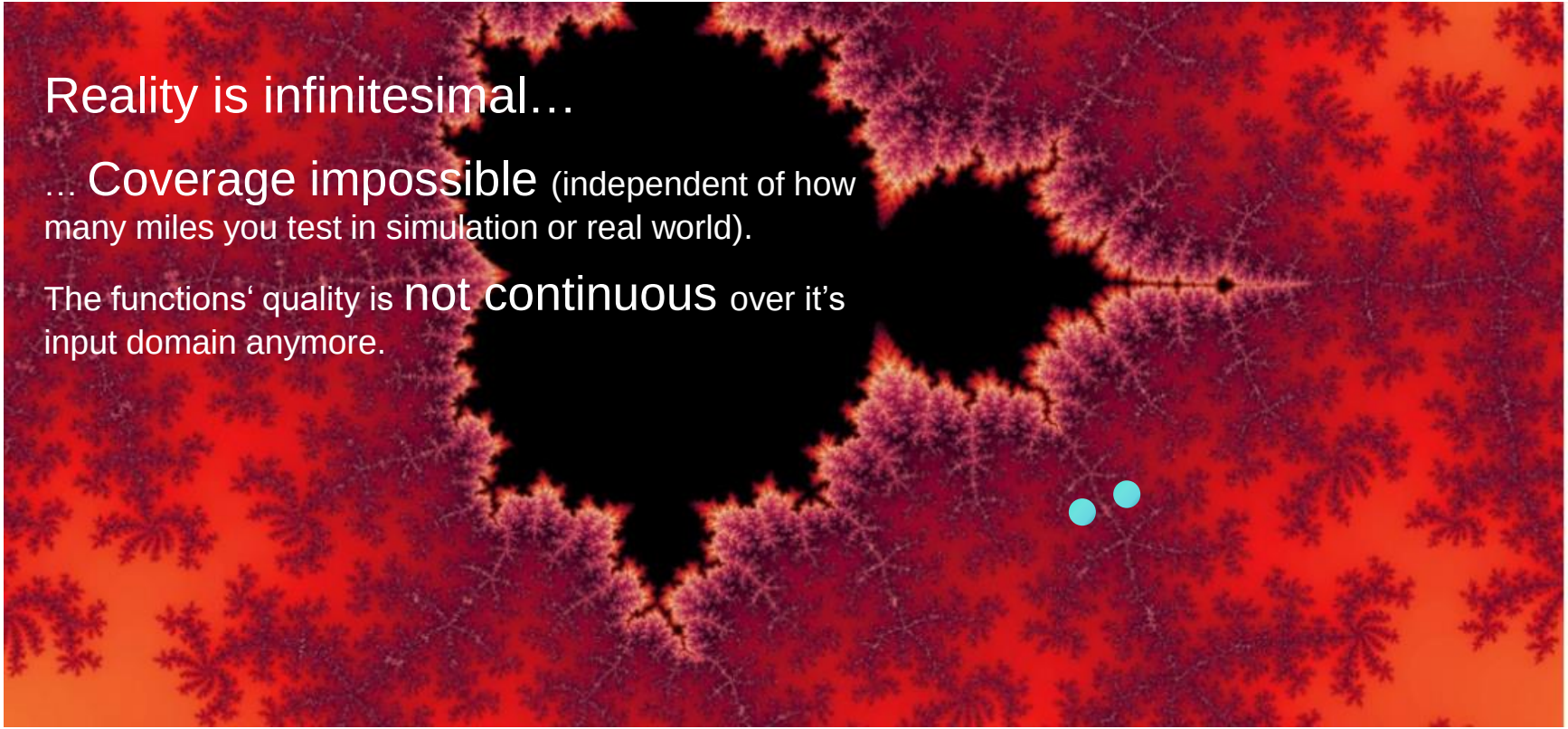
Dependency on training data

- Enormous **influence of quality** (Sample must represent situation / object to be recognized in all its facets)
- **Misdirected Training** (incorrect characteristics learned; overfitting)
- **Discontinuous quality function** (remains unpredictable independent of test intensity)
- **Advanced types of attacks** (subtle disturbance patterns)



[source] Sitawarin, Chawin, et al. "DARTS: Deceiving Autonomous Cars with Toxic Signs." *arXiv preprint arXiv:1802.06430* (2018).

CURRENT SAFETY-ASSURANCE APPROACHES DO NOT WORK FOR AI/ML

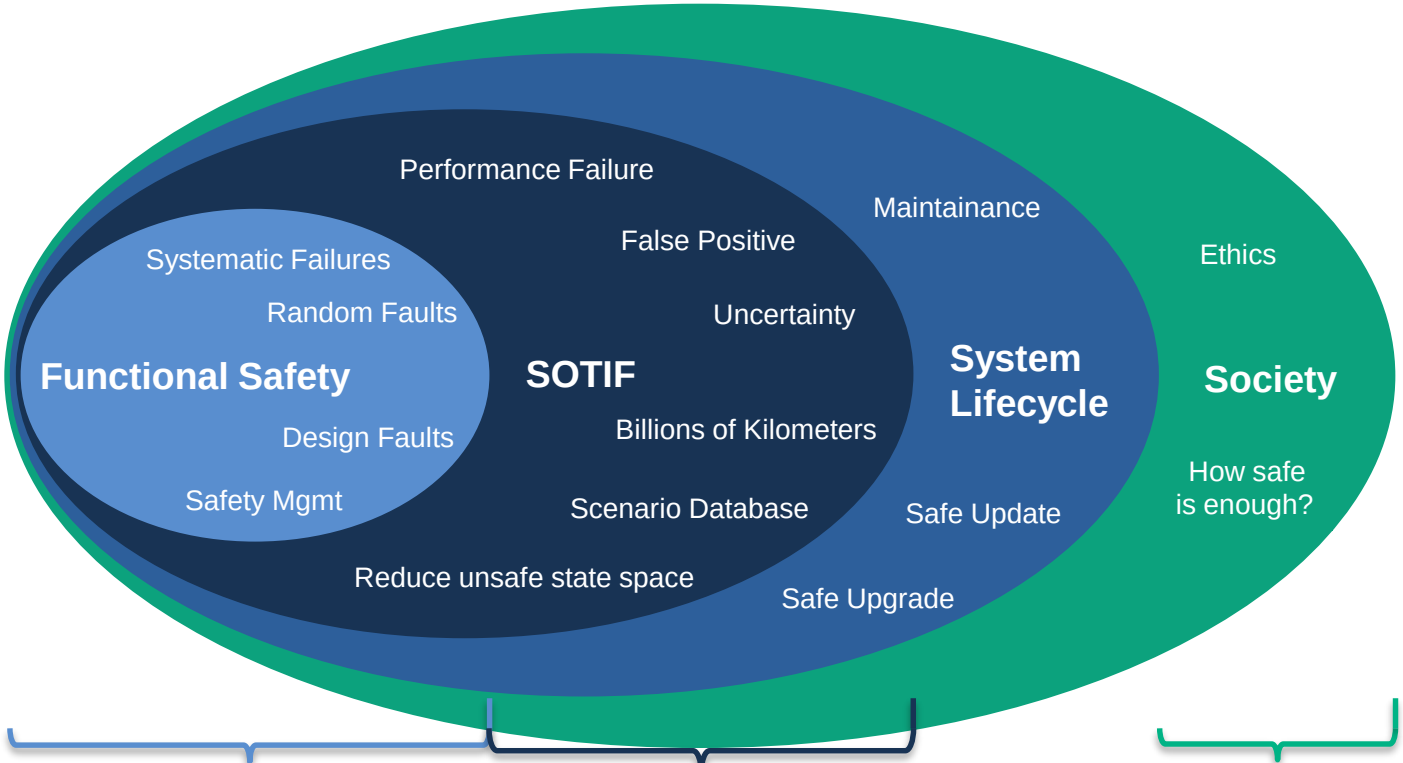


Reality is infinitesimal…

… Coverage impossible (independent of how many miles you test in simulation or real world).

The functions‘ quality is not continuous over it’s input domain anymore.

SAFETY STANDARDIZATION



ISO 26262 2nd edition (automotive),
EN 16590 / ISO 25119 (agriculture + forestry),
EN ISO 19014 (earthmoving + mining),
EN ISO 10218 (industry robots), …

ISO 21448 (SOTIF)

UL4600 (Draft Version)
„Safety First“ -> ISO TR (2020)

IEEE P700X

TWO CURRENT APPROACHES FOR ASSURING AI

01 **Improving the characteristics** of AI-based algorithms for enabling a sound assurance.

02 Assuming AI as **untrustworthy** and using **intelligent safety mechanisms** for assuring the systems' resilience even if the AI-algorithm fails.

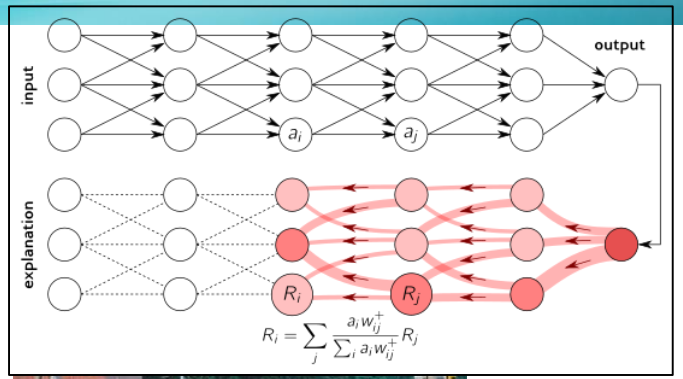
(1) ASSURING ARTIFICIAL INTELLIGENCE

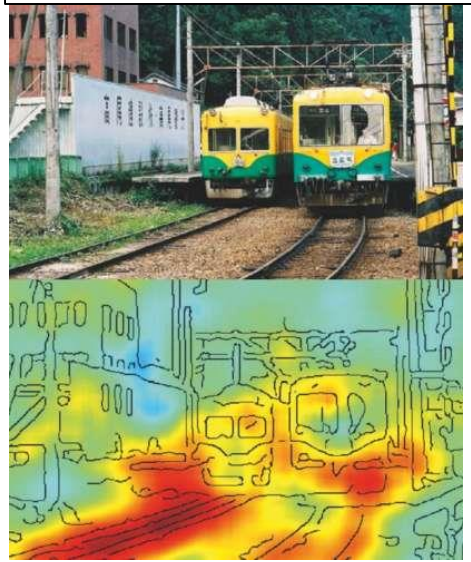
Assurable AI, by improving

- **Explainability:** Traceability of how the algorithm came to a concrete result (for one specific case)
- **Transparency**: the algorithms behavior is reproducible, in general (for any thinkable case)
- **Robustness**: small modifications of the algorithm's input lead to only a small impact on its output

Current status of assuring AI-based algorithms

- very first steps, e.g. using **saliency maps** by Layer-wise Relevance Propagation (LRP)
- AI should be considered as **untrustworthy** from a resilience point of view





[Source] http://www.heatmapping.org/

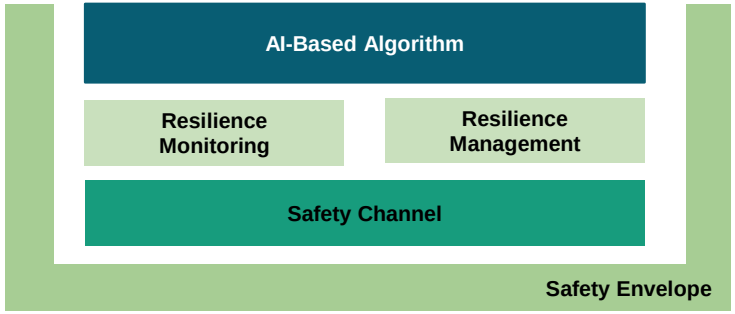
(2) SAFEGUARDING AI USING SAFETY ENVELOPES

AI-based algorithm

- is considered **untrustworthy**
- has **no direct access** to its environment
- its access to actuators is prevented and overwritten by a **safe fallback-function**, in case that the AI exposes an implausible / unsafe behavior

Safety Envelope

- The concept as such is established to safeguard **untrustworthy components** within a safety-critical system.



SAFETY-PATH ARCHITECTURE

Safety Path

Safety World Model

Trajectory Monitor

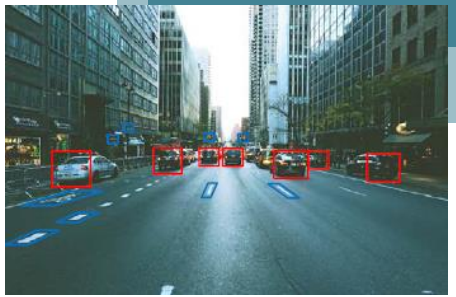
Act

Sensors

Perception

Plan

Performance Path



PUT A WILD HORSE INTO A CAGE?

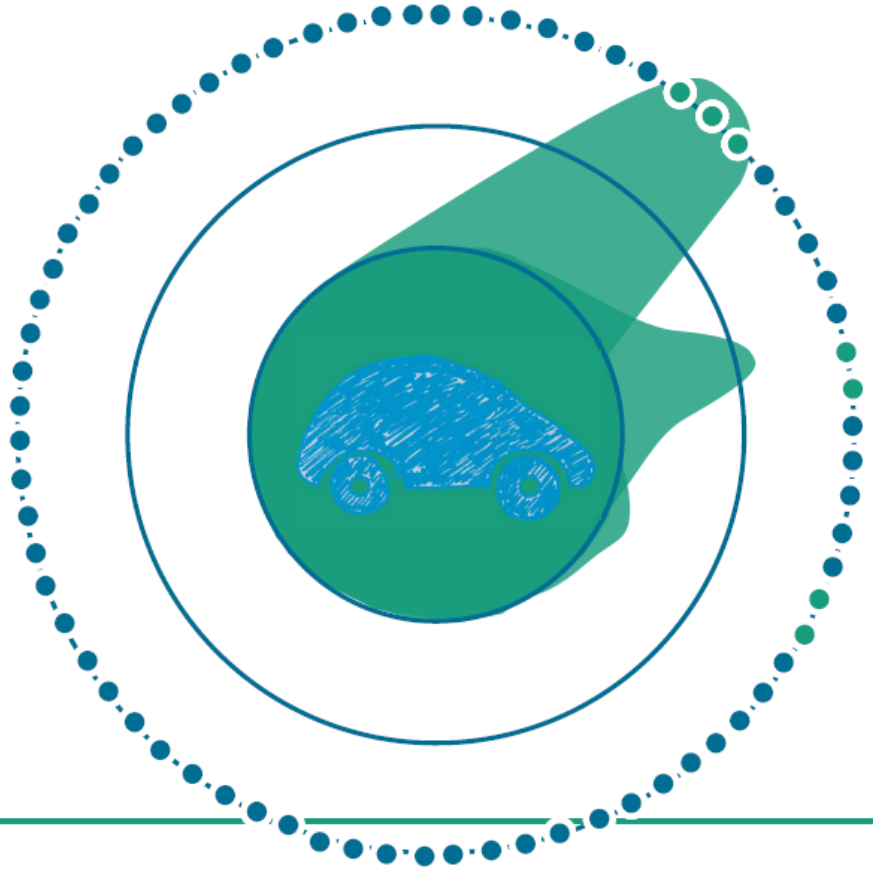
Simplification conflict…

…Safety-Cages / Envelopes are often too restrictive and thus impeding the benefits of AI/ML.

How to define safeguarding mechanisms in spite of so many uncertainties?

[source] HippoSport GmbH

ADAPTIVE SAFETY SPACE



Dynamic Is-Situation

Dynamic Safety-Space ➔ Adaptive Safety-Space

φ - n-dimensional context classification space

→ || || - degree of freedom

ADAPTIVE SAFETY MANAGEMENT - SAFETY ASSURANCE @

Information

External Context

Monitoring

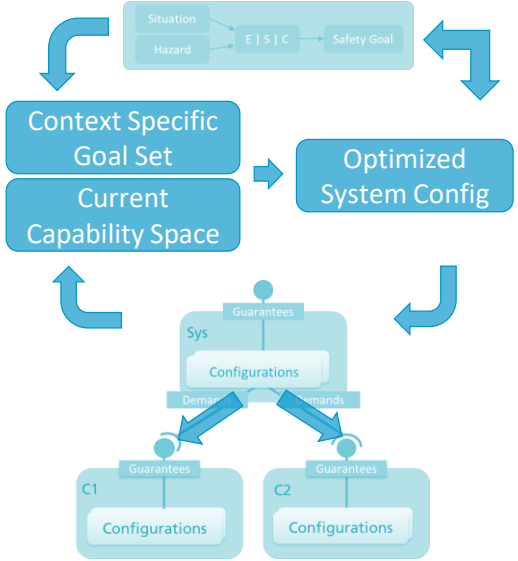
Context awareness

Self-Awareness

Knowledge

Safety Model @ Runtime

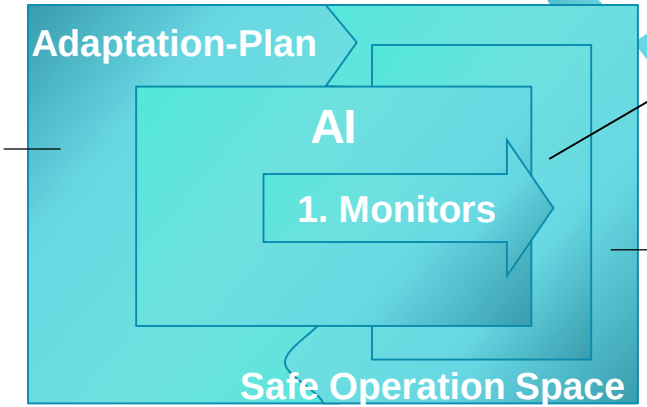
Conscious Management



EXAMPLE: 4-LAYER SAFETY ARCHITECTURE

1. Layer: White-Box / Gray-Box AI-Monitors (Operational)
2. Layer: Safety-Cage (Envelopes) (Operational)
3. Layer: Adaptive Safety Management (Tactical)
4. Layer: Continuous Learning (Strategic)

4. Continuous Safety Management Platform
Learning from Field Data/ Incidents / Accidents



3b. Adaptive System Manager
- Performance Monitors
- Self-Optimization
- Adaptation Manager

2. Safety Cage
- (Dedicated AI-) Monitors
- Sandboxing

3a. Dynamic Safety Manager
- Dynamic Risk Assessment
- Dynamic Capability Assessment

OUR SCIENTIFIC COMPETENCIES

Resilient Artificial Intelligence
Robust, explainable, transparent AI
for resource-limited systems

Adaptive and Adaptable Architectures
Using most modern software technologies for safety-critical
and highly-reliable technical systems

Safety, Reliability, Availability
Assuring safety, reliability, and availability in spite of and
because of using modern software technology

SAFE INTELLIGENCE

THANK YOU FOR YOUR TIME AND ATTENTION!

Dominique Seydel, M.Sc.

Research Engineer
Fraunhofer Institute for Embedded Systems and Communication Technology ESK

Hansastraße 32 I 80686 Munich
dominique.seydel@esk.fraunhofer.de I Tel. +49 89 547088-363