Fraunhofer
IKS

Fraunhofer-Institut für Kognitive
Systeme IKS

September 9, 2025

Resilience Revisited –
Assuring Safety in the Face of the Unpredictable

Mario Trapp

Generated by AI (Adobe PS)

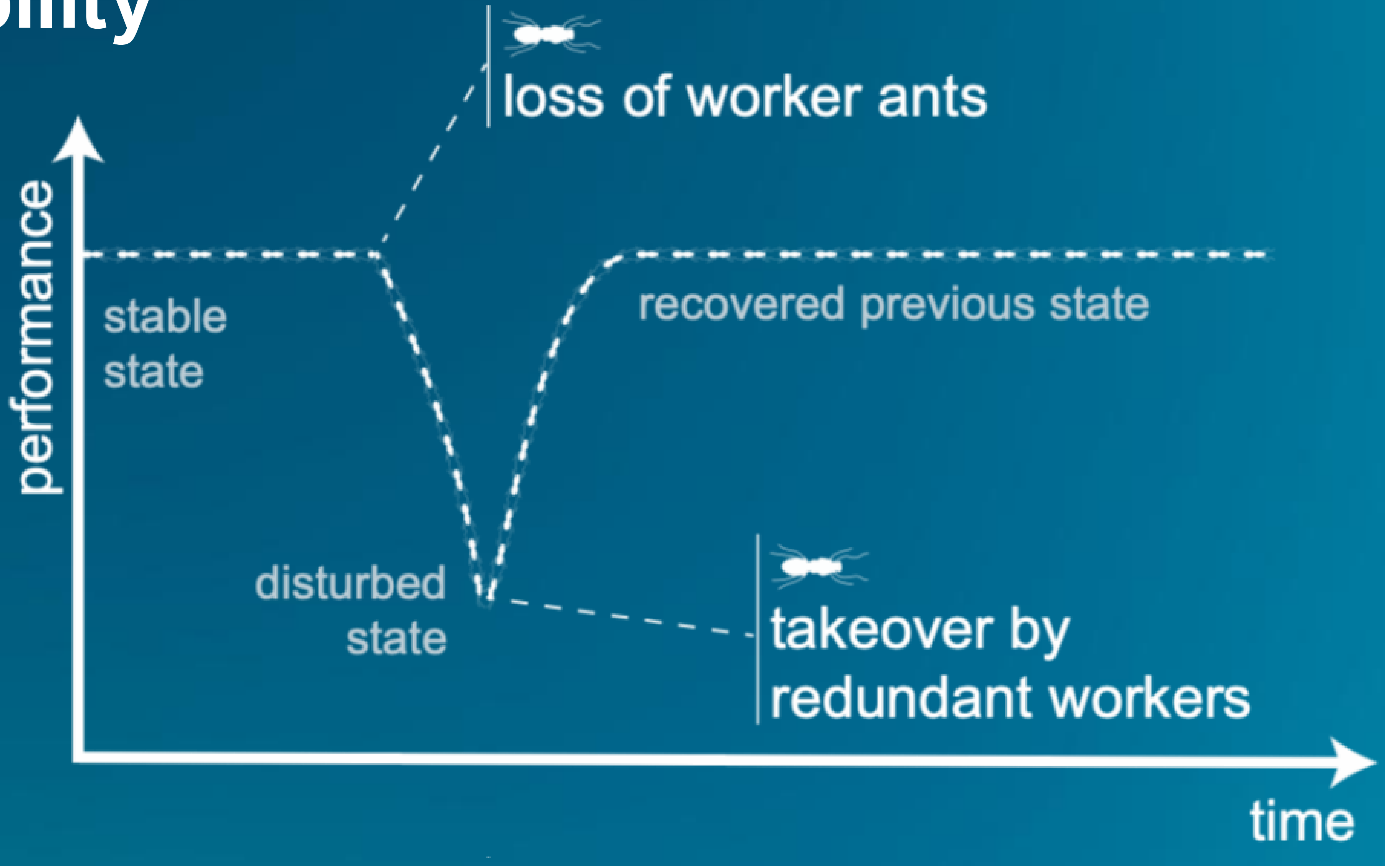
Resilience

Persistence of dependability when facing changes.

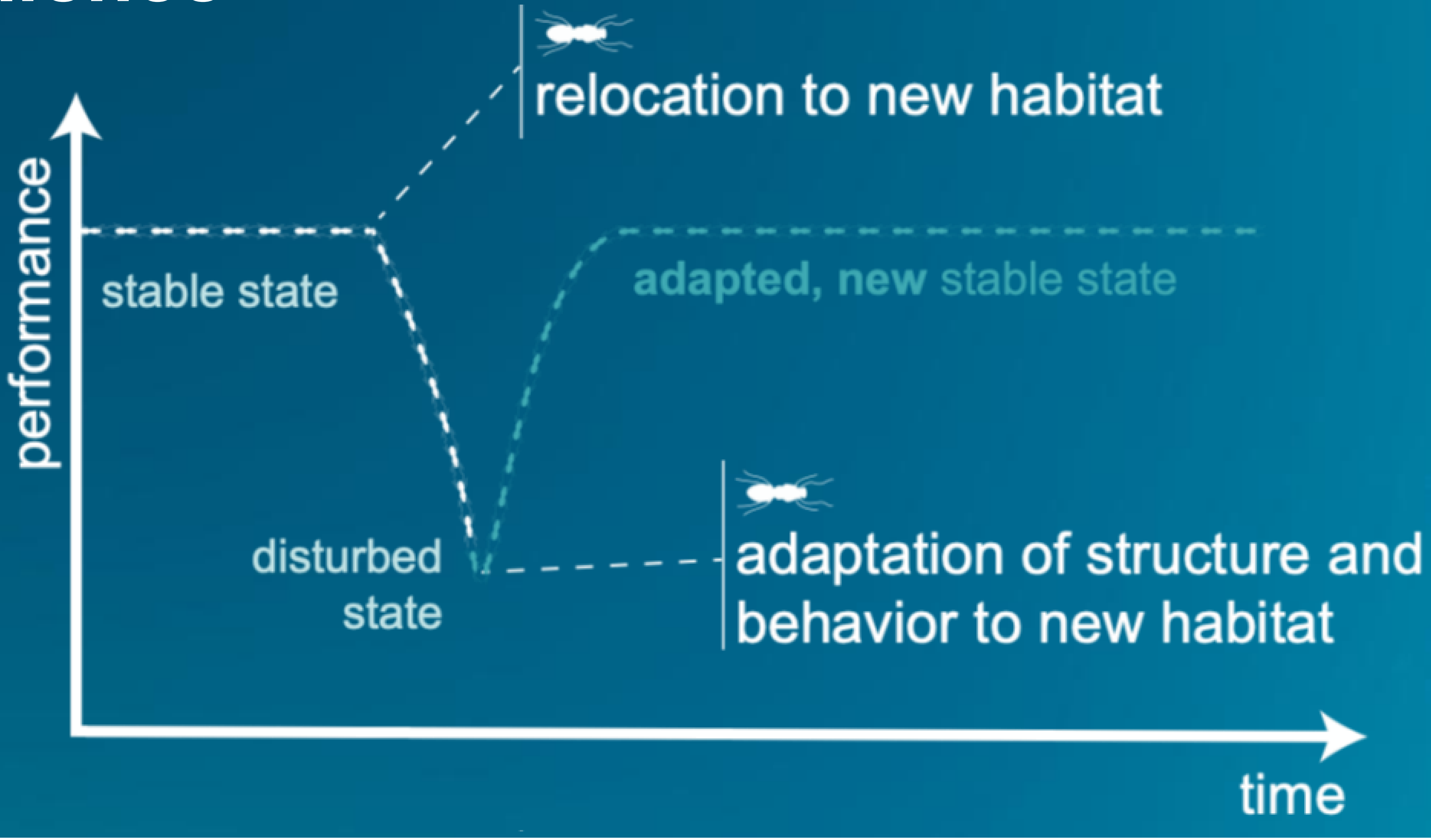
[Laprie]

Holling, Crawford S. “Resilience and Stability of Ecological Systems.” Annual Review of Ecology and Systematics 4, no. 1 (1973): 1–23.

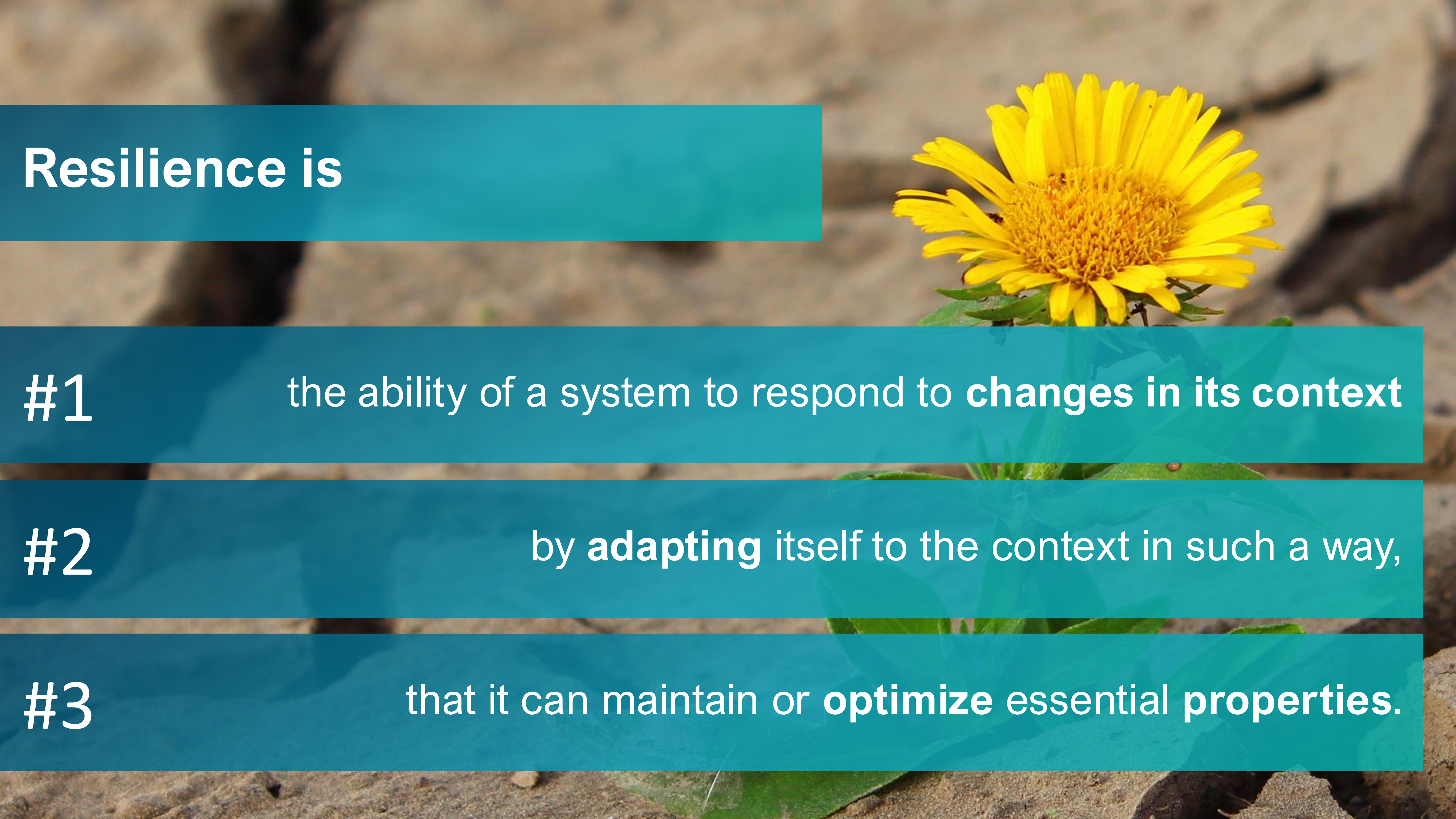
Stability


loss of worker ants
performance
stable state
recovered previous state
disturbed state
takeover by redundant workers
time

Resilience



relocation to new habitat
performance
stable state
adapted, new stable state
disturbed state
adaptation of structure and behavior to new habitat
time



Resilience is

#1 the ability of a system to respond to **changes in its context**

#2 by **adapting** itself to the context in such a way,

#3 that it can maintain or **optimize** essential **properties**.

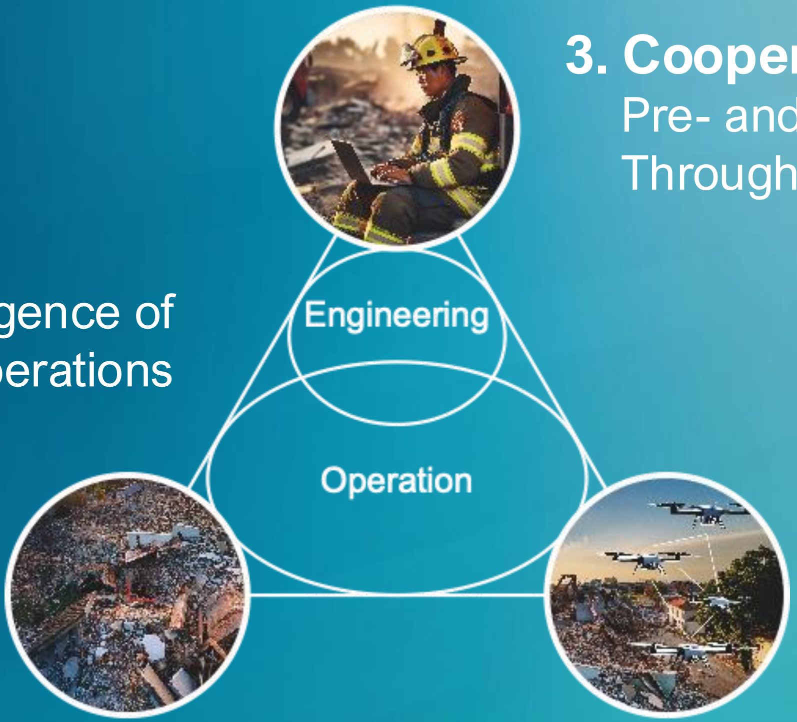
Resilience

Optimizing Utility whilst **Preserving Safety** in **Uncertain Contexts.**

Triale Intelligenz – Three-Fold-Intelligence

An Overview on our Research

TRION



3. Cooperative Intelligence
Pre- and In-Mission Adaptation
Through **Human-AI-Cooperation**

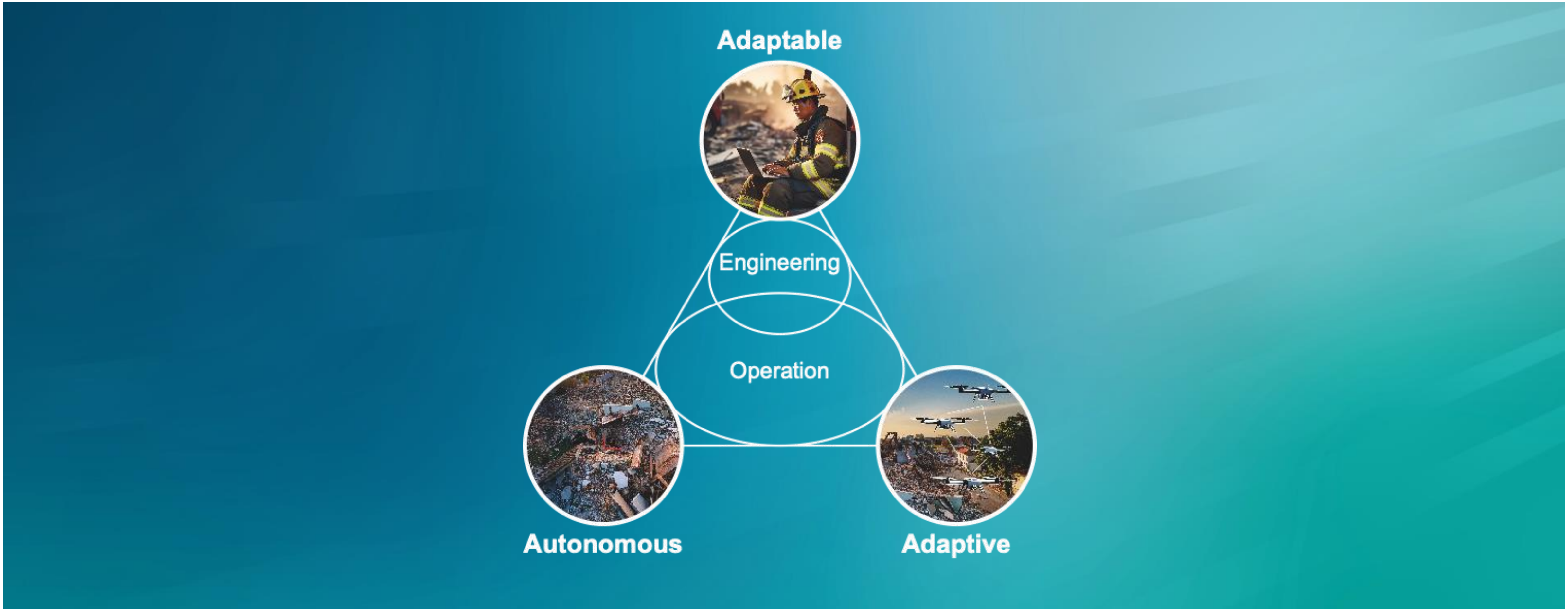
Seamless Convergence of
Engineering and Operations

1. Functional Intelligence
Autonomy through Resilient AI

2. Adaptive Intelligence
Continuous **Self-Adaptation**
to Context and System State

Overview

An Overview on our Research



Adaptable

Engineering

Operation

Autonomous

Adaptive

Overview

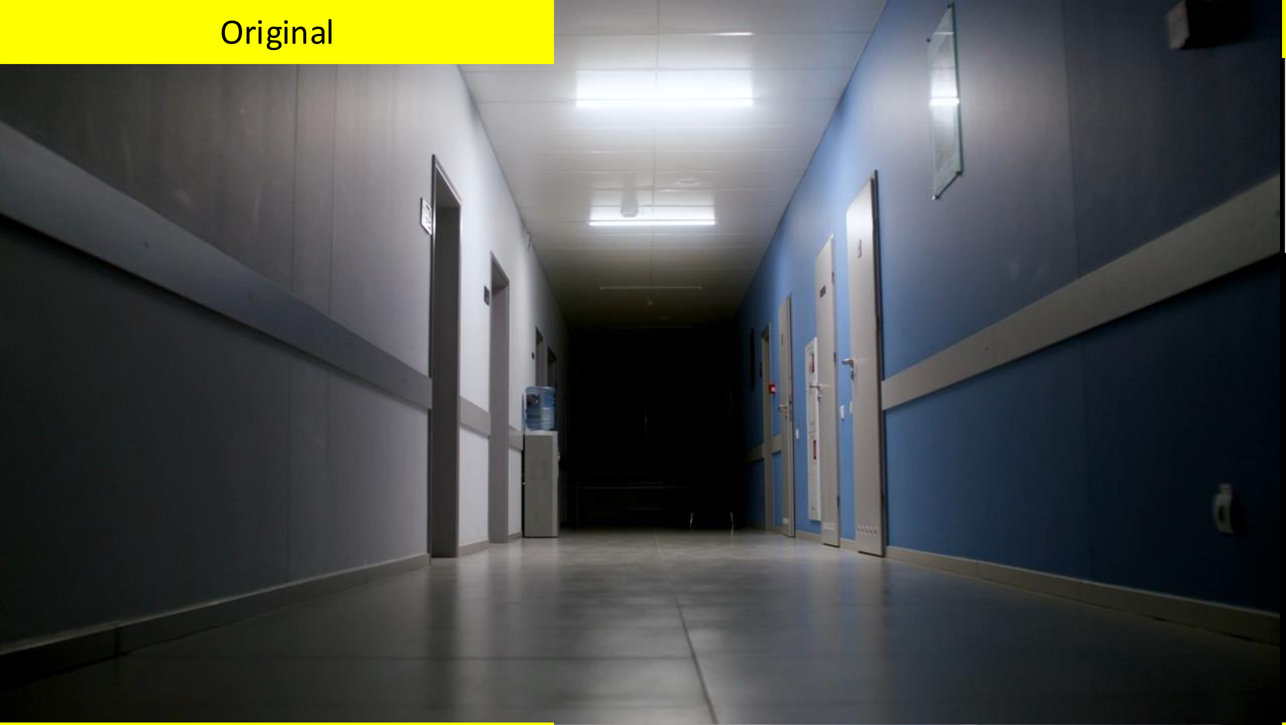
An Overview on our Research

Resilient AI

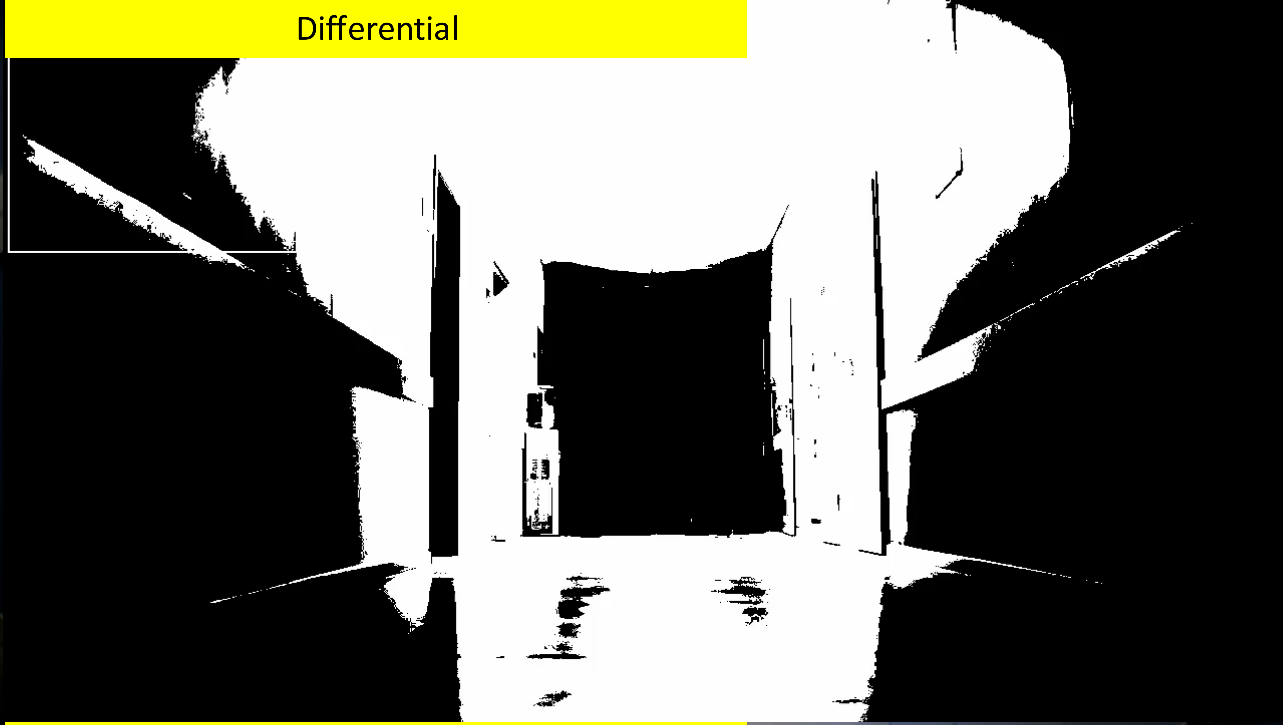
Resilient AI

1 It’s not about Safety Assurance but about Safe Design

Original



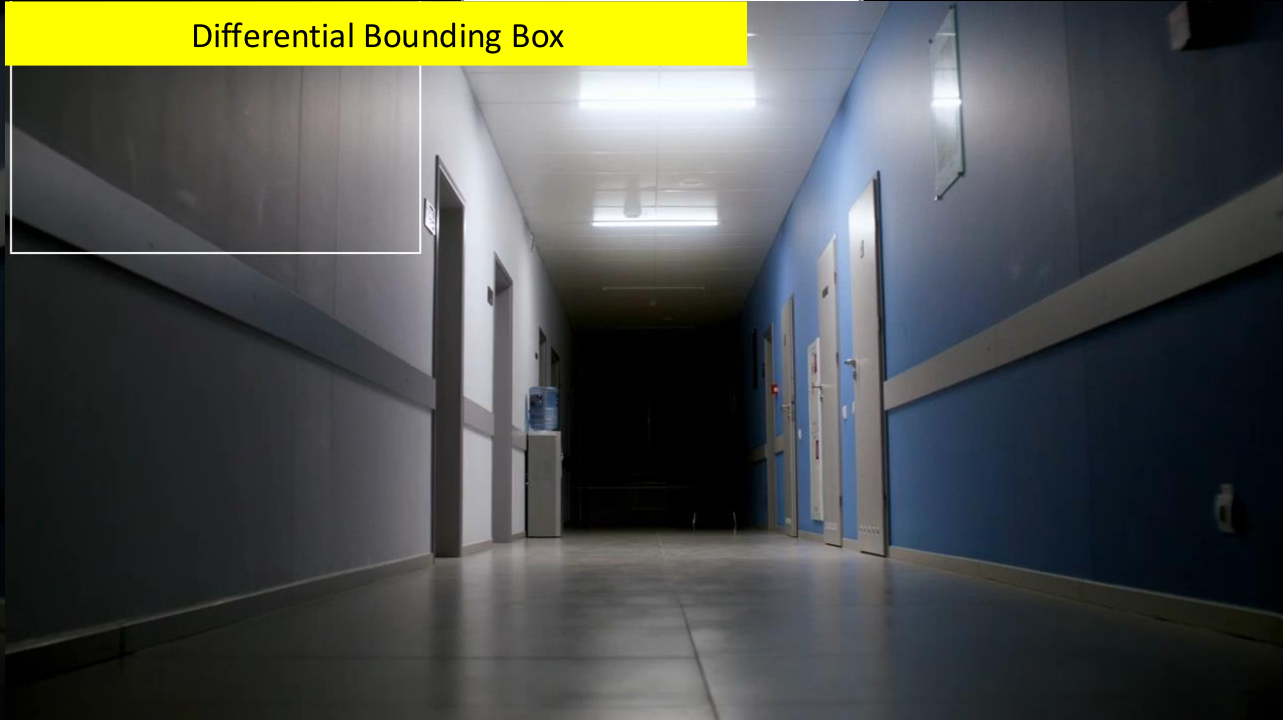
Differential



Yolo v4 (COCO)

2025/09/09

Differential Bounding Box



Example – Safe Person Detection

Resilient AI

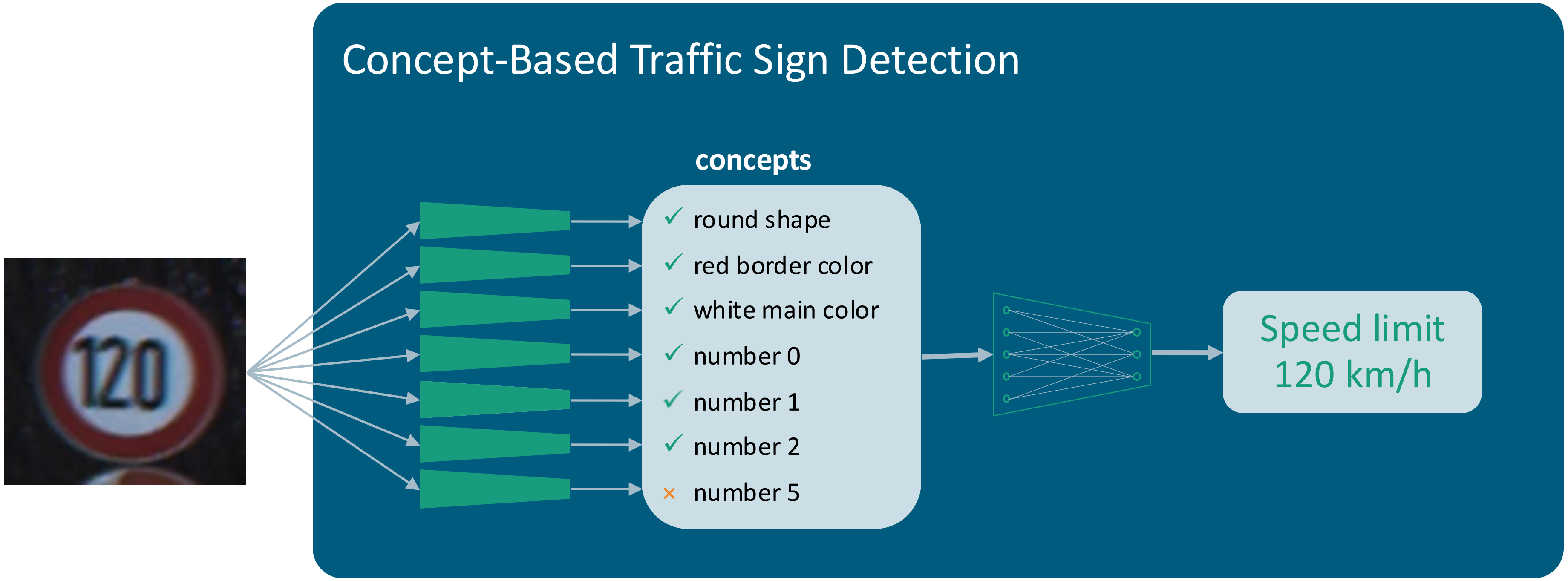
1 It’s not about Safety Assurance but about **Safe Design**

2 It’s not about Safe AI but about **Safe Systems**

3 It’s not about Trends but about **Value**

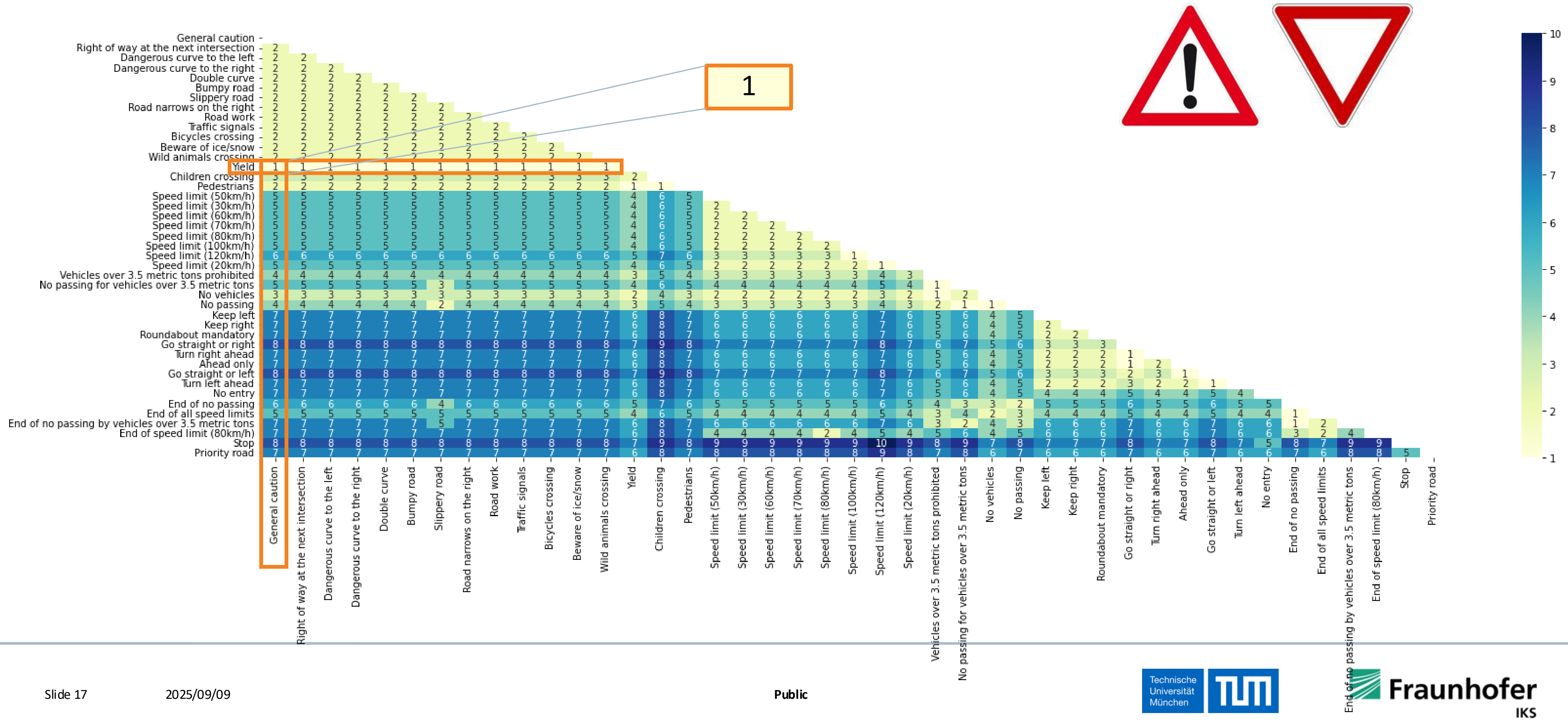
Example – Concept-Learning

Micro-Level Safety Architectures



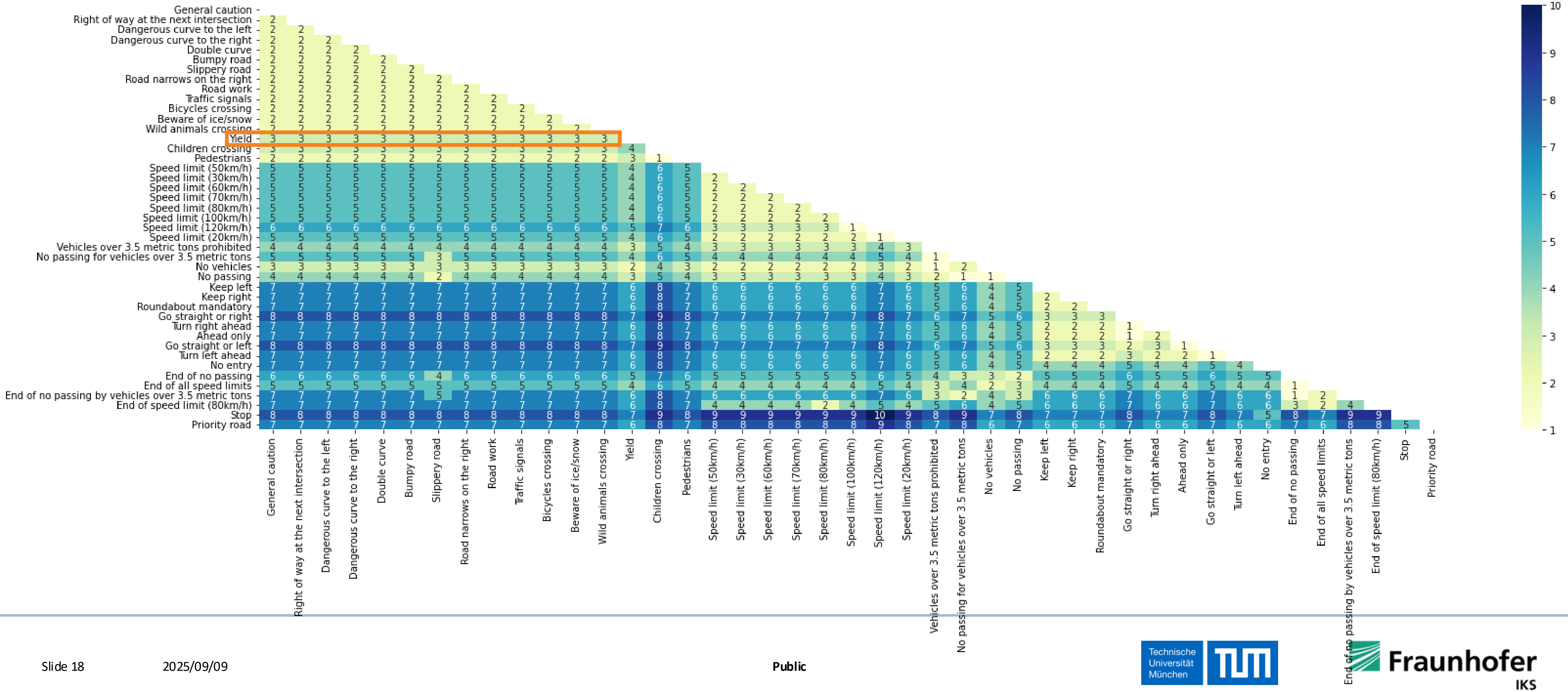
Micro-Level Safety-Analysis

Measuring Class Distance



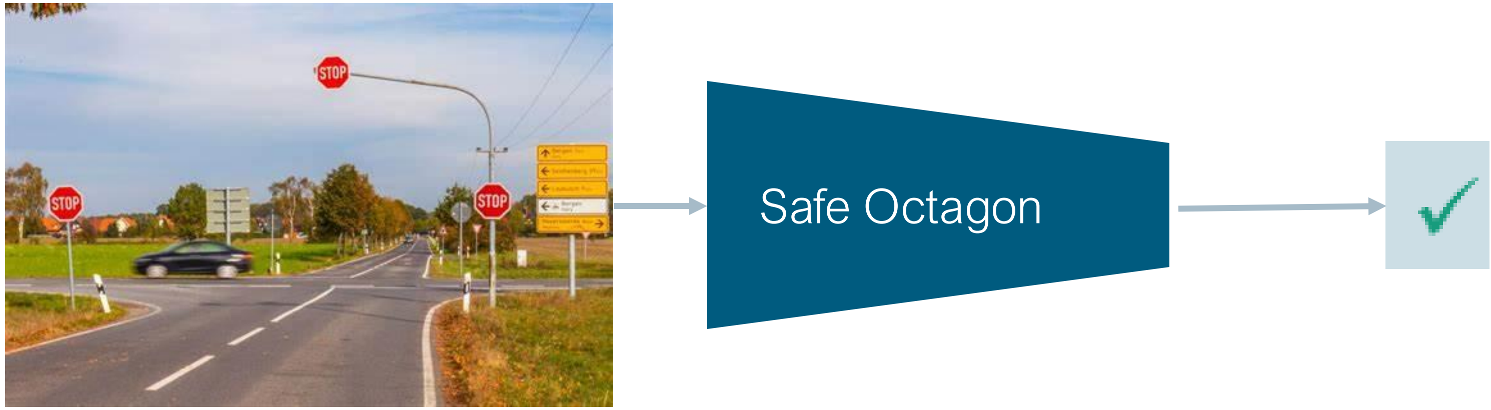
Micro-Measures

Example – Modifying Concept Architecture: Upward Triangle + Downward Triangle



Micro-Level Safety-Architectures

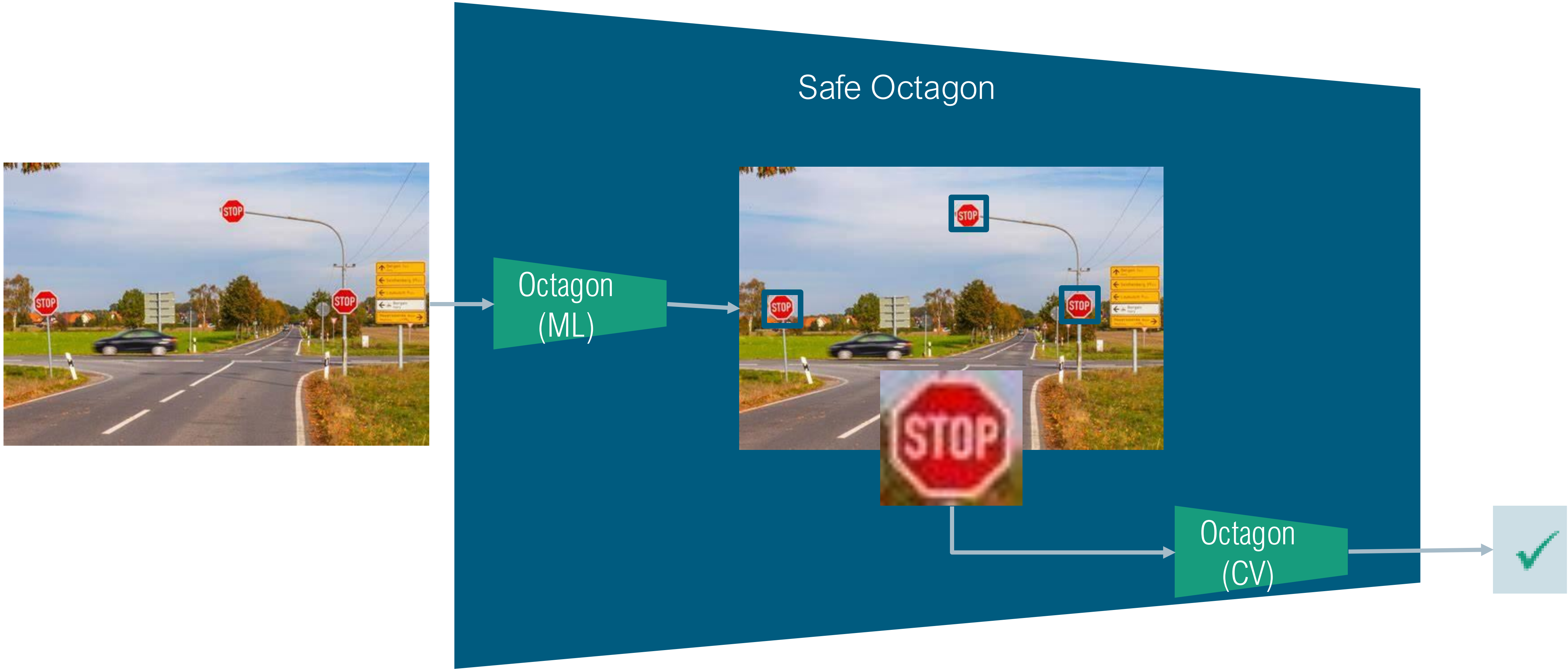
Concept-Level Counter-Measures



Safe Octagon

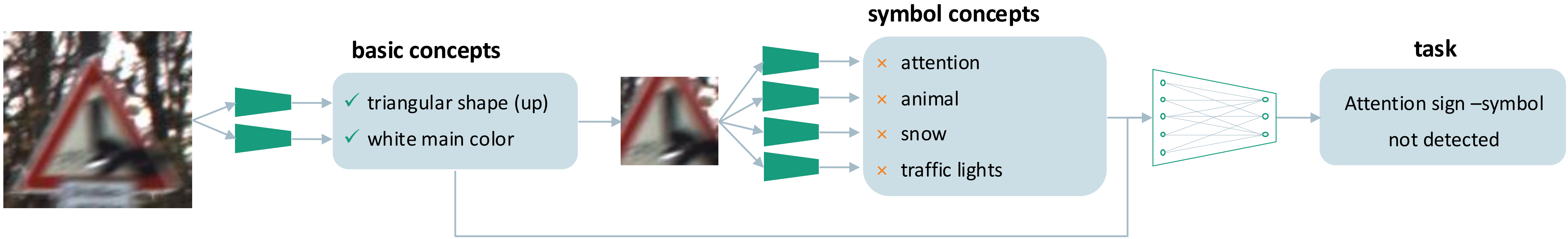
Micro-Level Safety-Architectures

Concept-Level Counter-Measures



Hierarchical Classification

Concept Chains



basic concepts

✓ triangular shape (up)

✓ white main color

symbol concepts

× attention

× animal

× snow

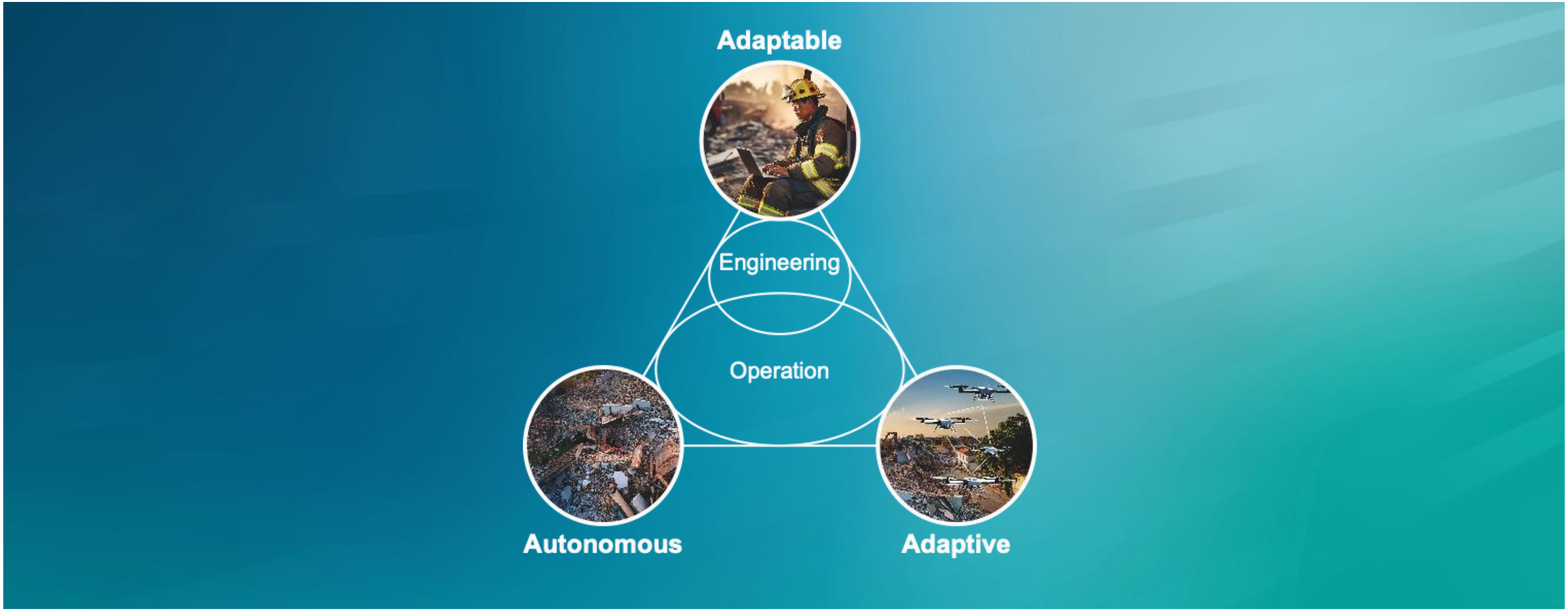
× traffic lights

task

Attention sign –symbol not detected

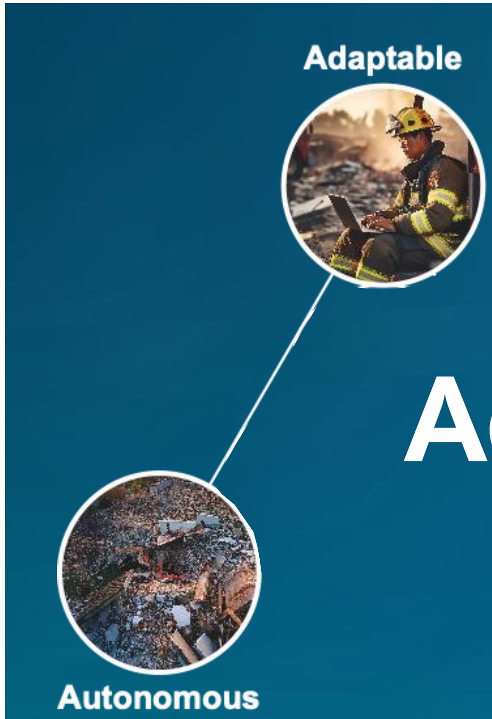
Overview

An Overview on our Research



Overview

An Overview on our Research

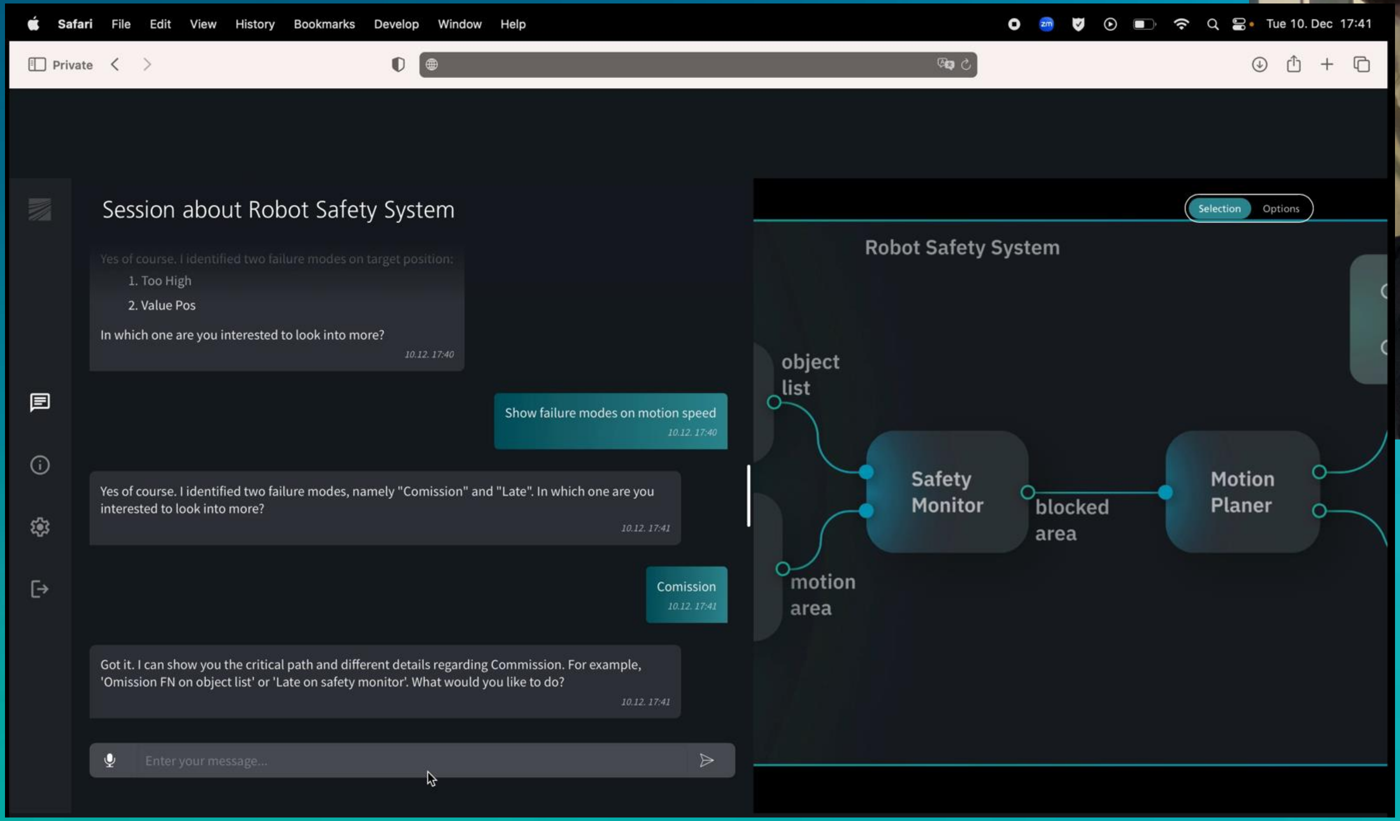


Adaptable Autonomy

∞

Cycle-Time Automotive: 10-14 days

AI-Assisted Safety Engineering



Session about Robot Safety System

Yes of course. I identified two failure modes on target position:

1. Too High
2. Value Pos

In which one are you interested to look into more?
10.12. 17:40

Show failure modes on motion speed
10.12. 17:40

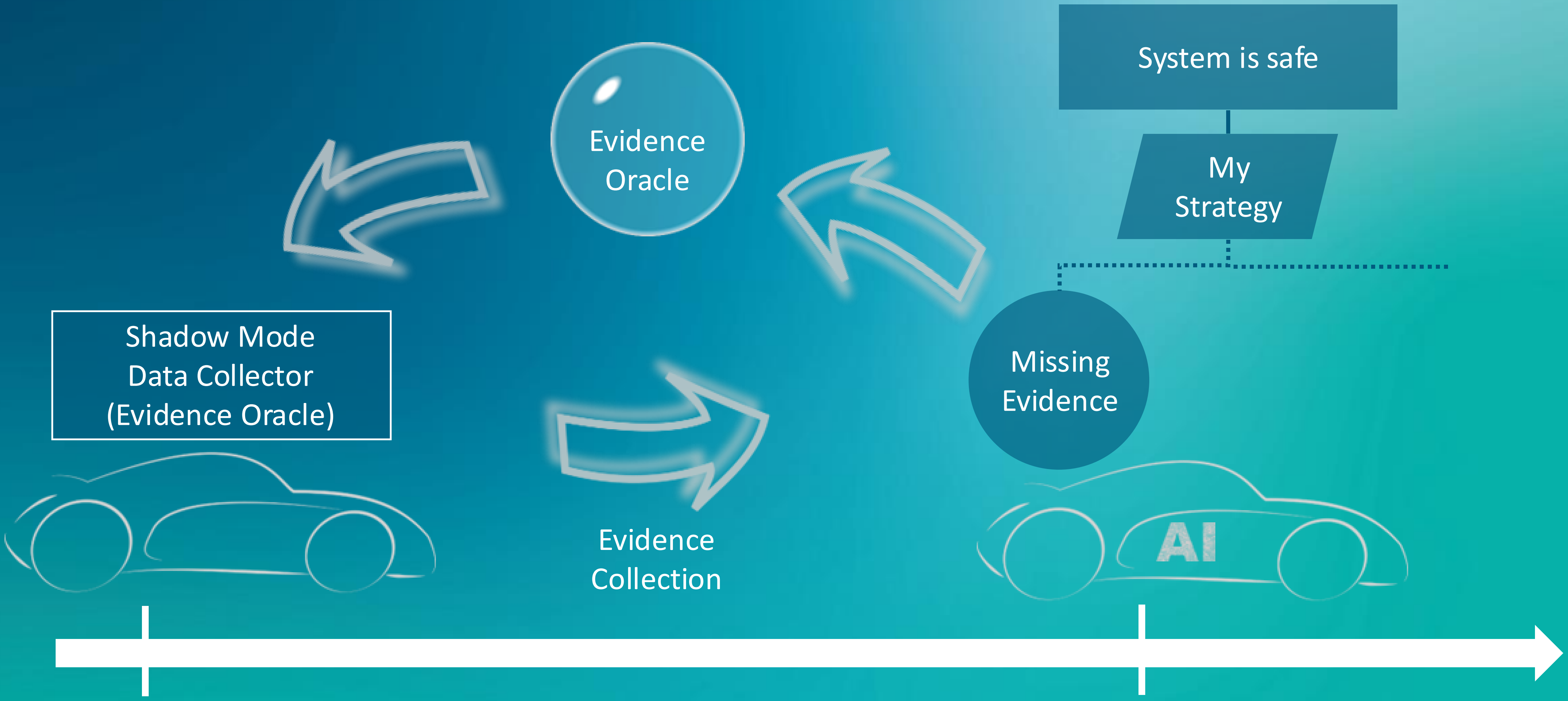
Yes of course. I identified two failure modes, namely "Comission" and "Late". In which one are you interested to look into more?
10.12. 17:41

Comission
10.12. 17:41

Got it. I can show you the critical path and different details regarding Commission. For example, 'Omission FN on object list' or 'Late on safety monitor'. What would you like to do?
10.12. 17:41

Session about Robot Safety System
Hello, you start a new session on Robot Safety System. What do you want to check more deeply today?
Enter your message...
Robot Safety System
object list
Safety Monitor
blocked area
Motion Planer
motion area
motion speed
target position

Predictive Assurance



System is safe

My Strategy

Missing Evidence

Evidence Oracle

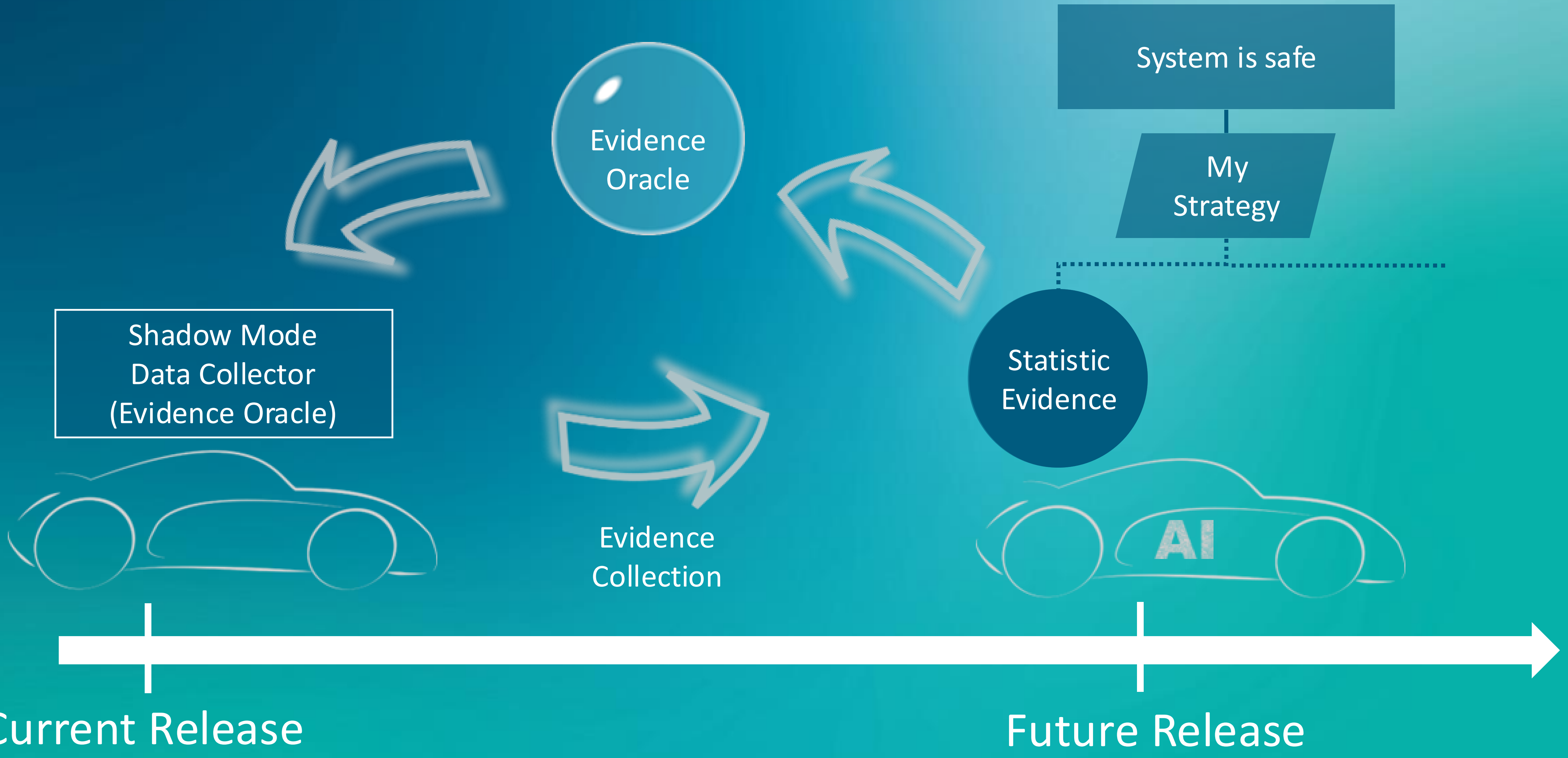
Shadow Mode Data Collector (Evidence Oracle)

Evidence Collection

Current Release

Future Release

Predictive Assurance


System is safe

My Strategy

Evidence Oracle

Statistic Evidence

Shadow Mode Data Collector (Evidence Oracle)

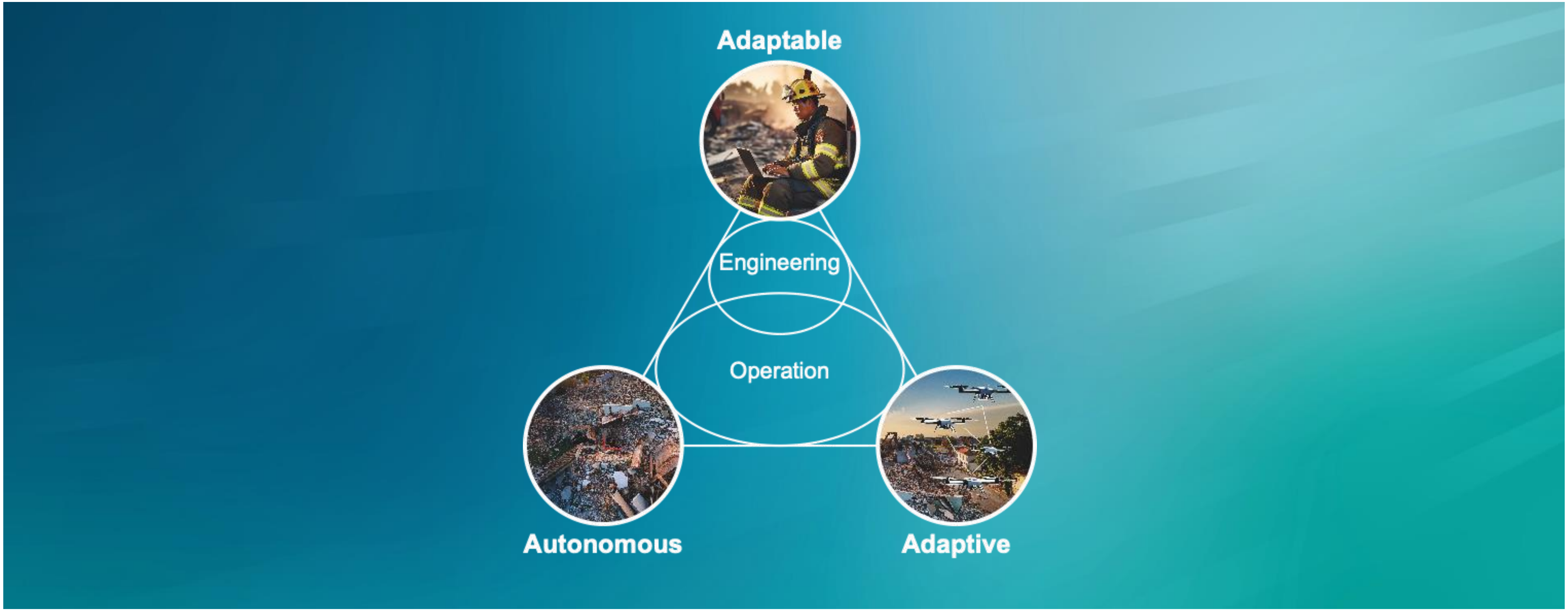
Evidence Collection

Current Release

Future Release

Overview

An Overview on our Research



Overview

An Overview on our Research

Adaptivity

Resilience

Optimizing Utility whilst Preserving Safety in Uncertain Contexts

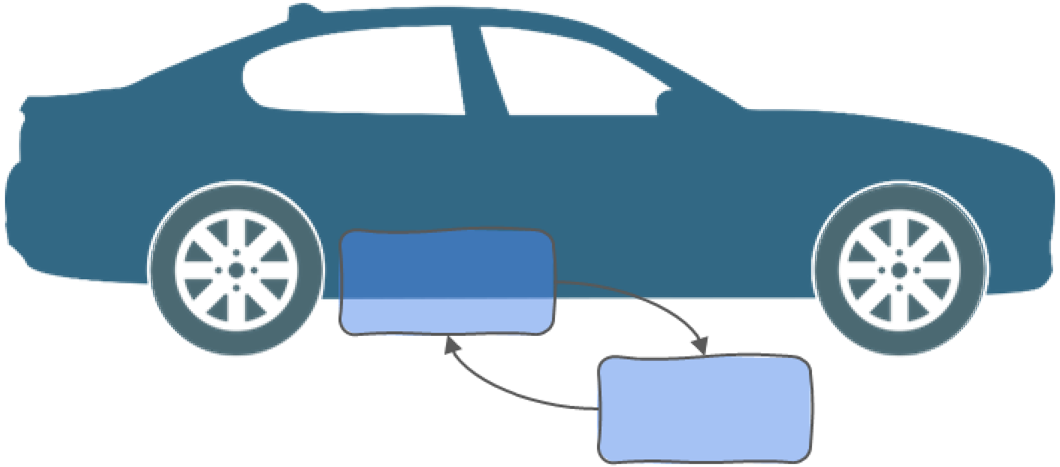
[Trapp]

Why I

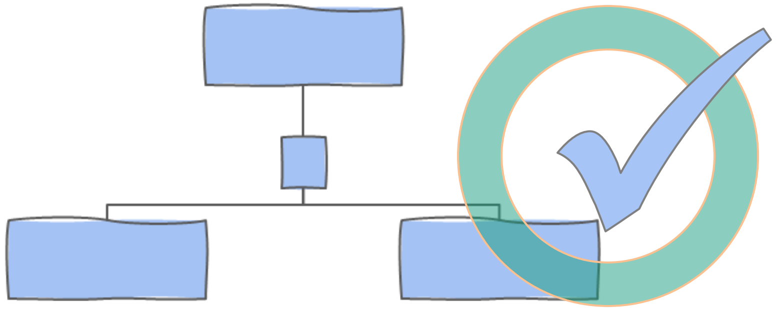
Why do we need resilience?

How to Engineer Resilience?

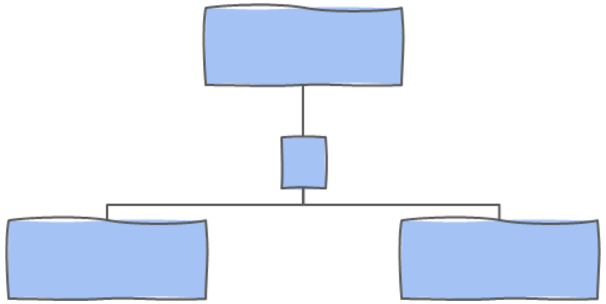
Safety Assurance Today

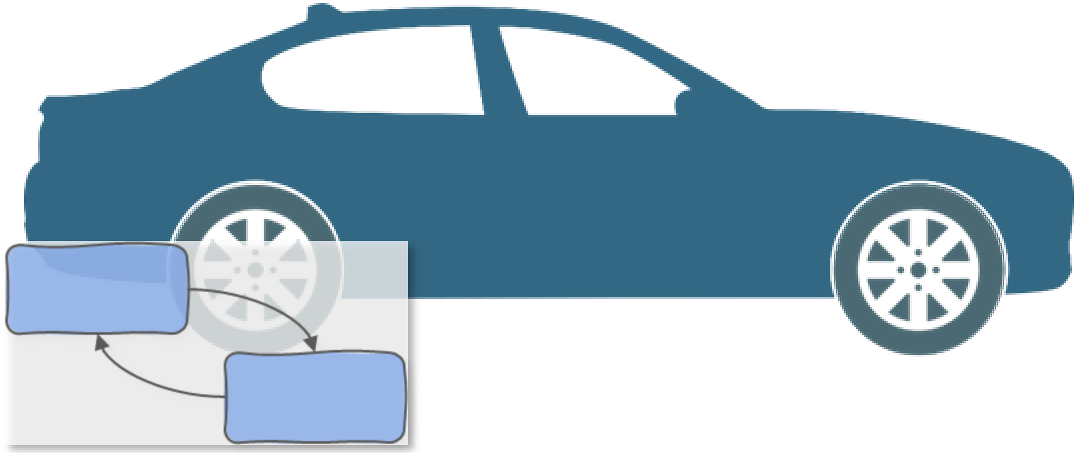


ANTICIPATE. ANALYZE. ASSURE.



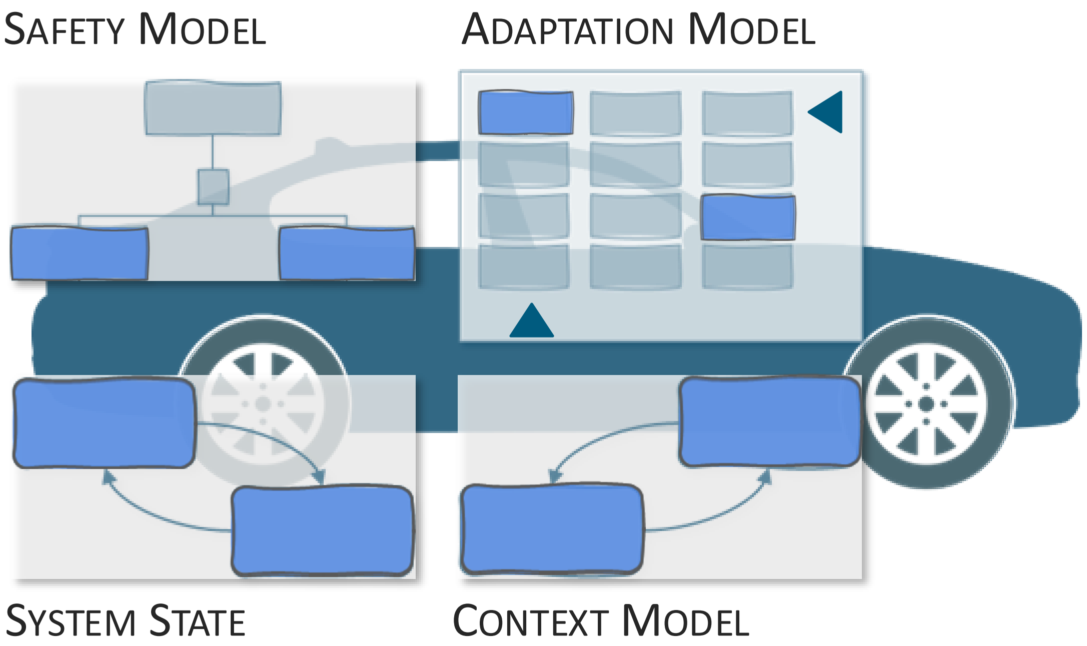
Adaptive Safety

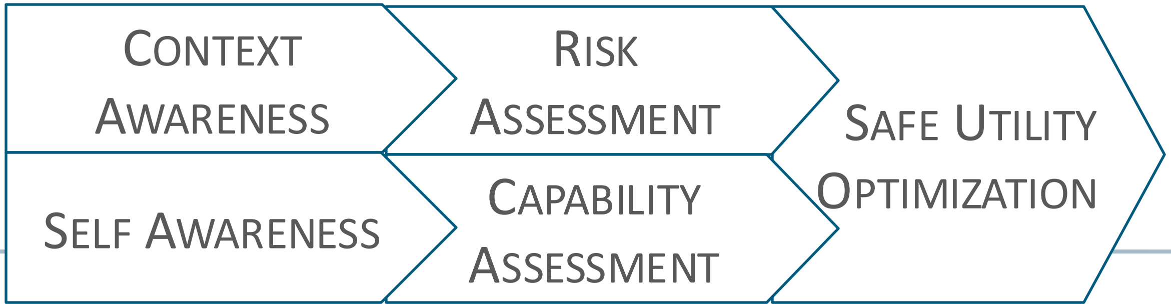




SYSTEM STATE

Adaptive Safety





CONTEXT AWARENESS

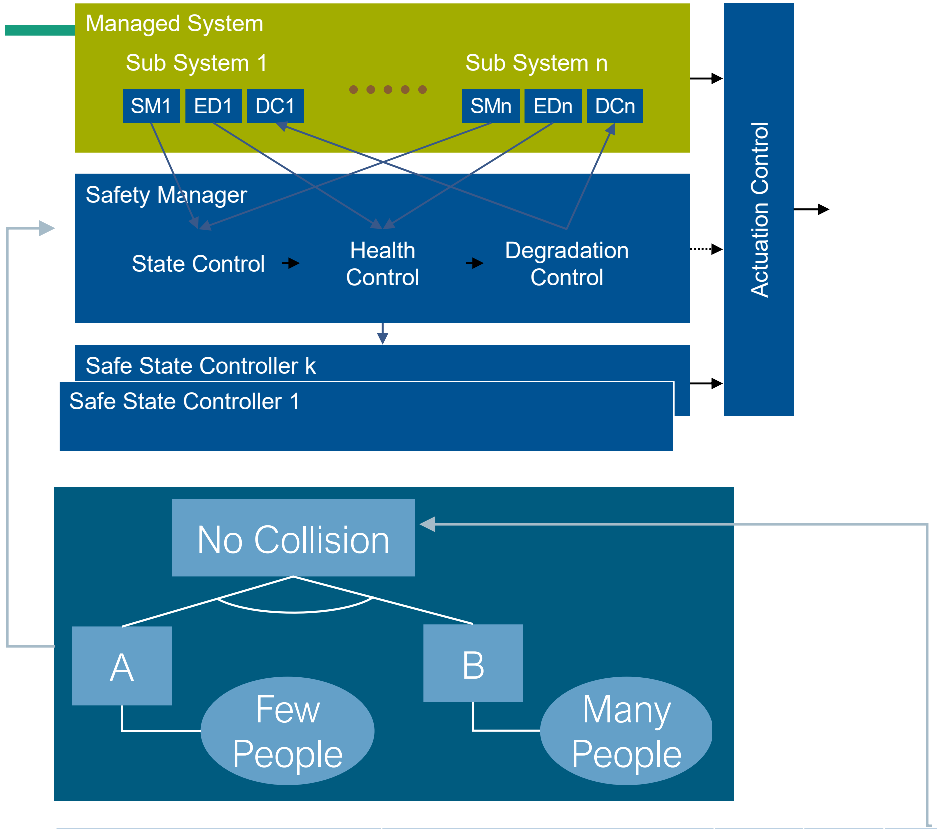
SELF AWARENESS

RISK ASSESSMENT

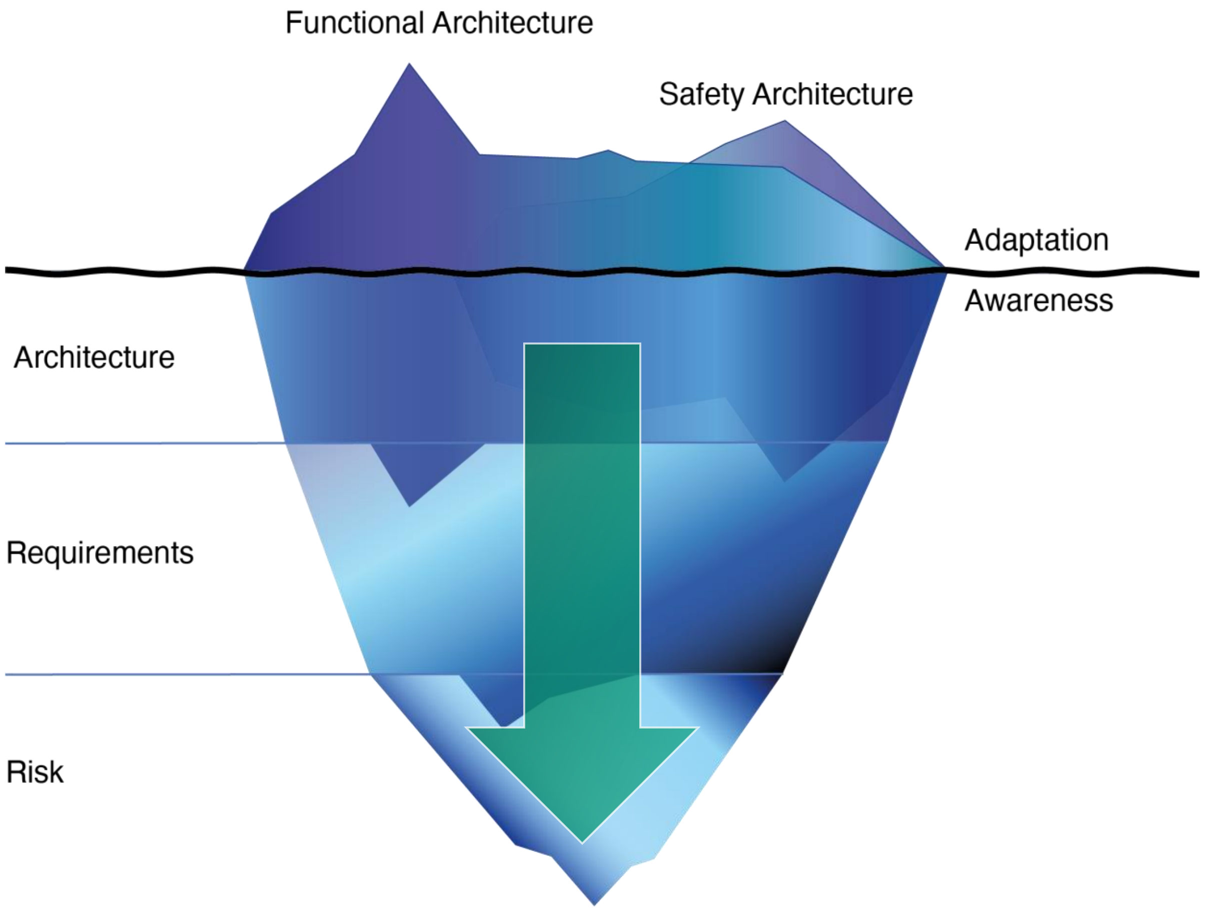
CAPABILITY ASSESSMENT

SAFE UTILITY OPTIMIZATION

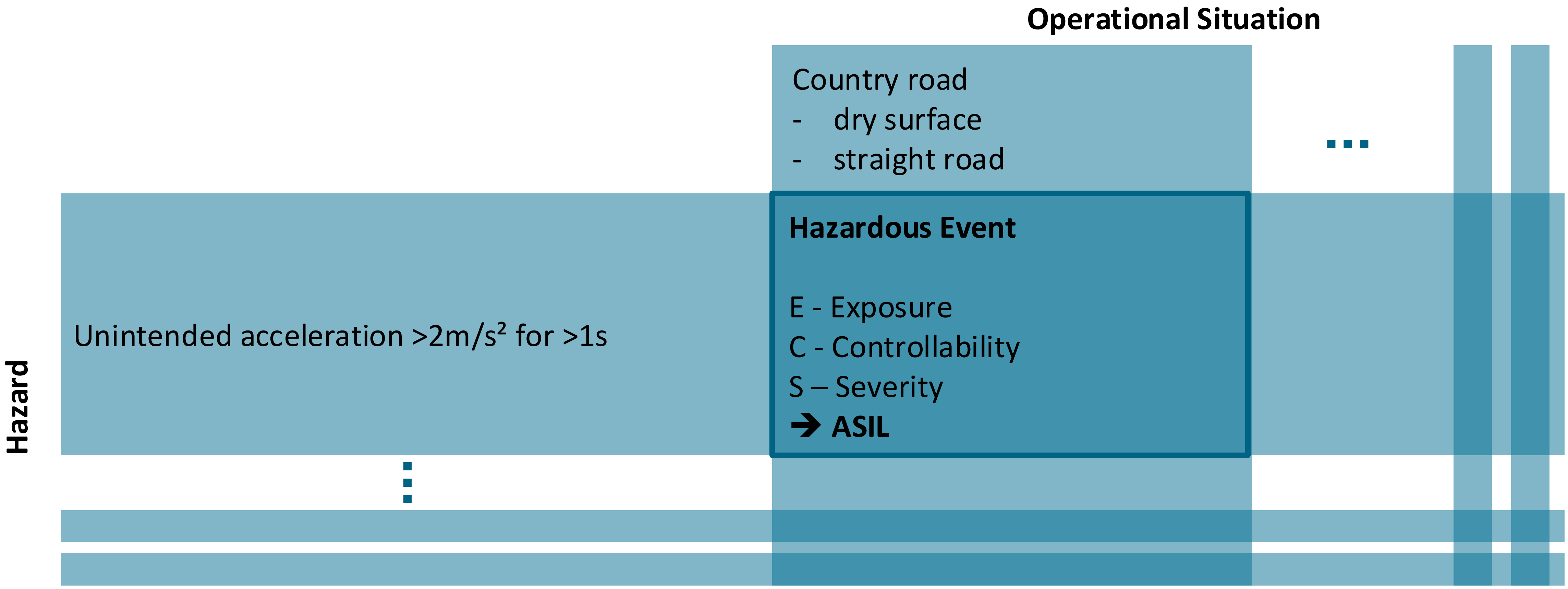
The Iceberg Model



Failure	Situation	C	F	P	W	SIL
does not stop despite obstacle	***Slow*** speed (≤ 1m/s) in ***narrow*** aisle in ***restricted*** area	C2	F1			-
does not stop despite obstacle	***Full*** speed (> 1m/s) in ***narrow*** aisle in ***restricted*** area	C3	F1	P2	W3	1



Example: Risk Models @ Runtime



Operational Situation

Country road
- dry surface
- straight road

…

Hazard

Unintended acceleration >2m/s² for >1s

⋮

Hazardous Event

E - Exposure
C - Controllability
S – Severity
➔ **ASIL**

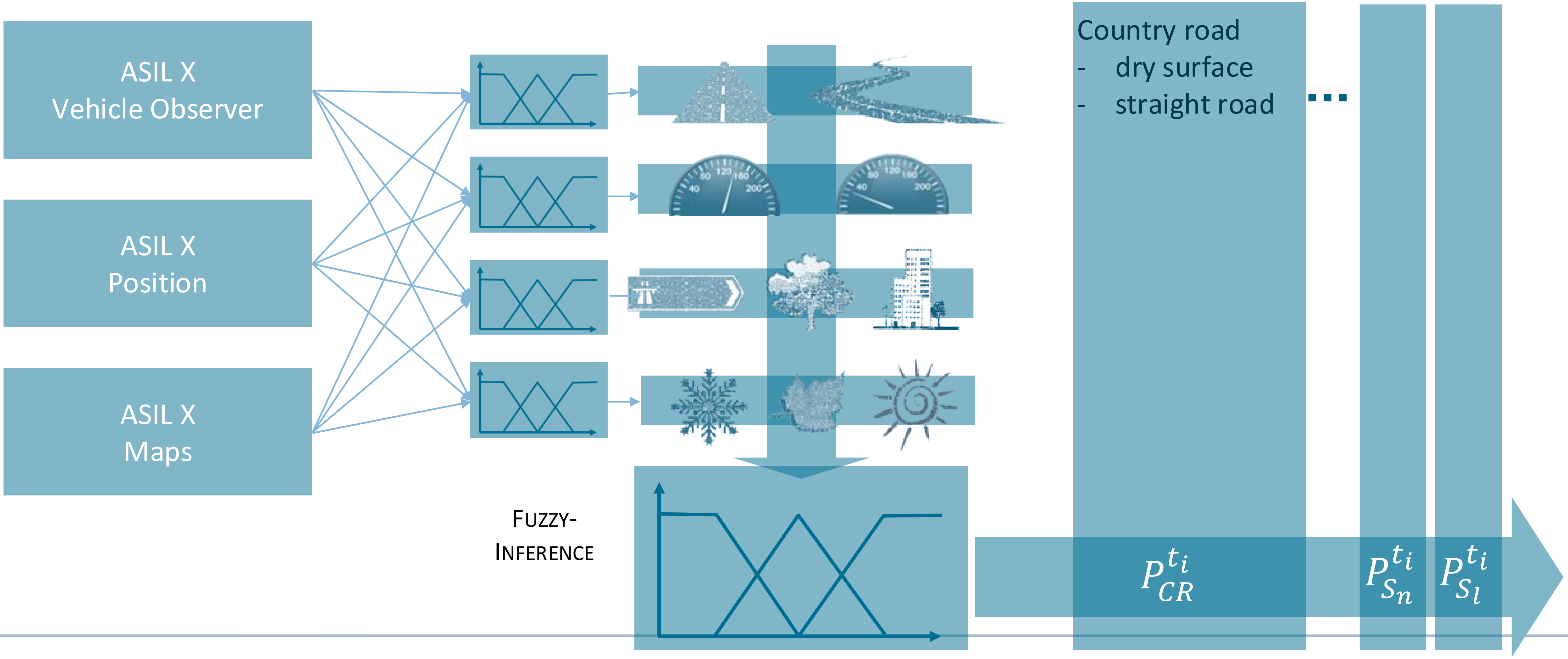
***ASIL: Automotive Safety Integrity Level [ISO26262]**

The design time model

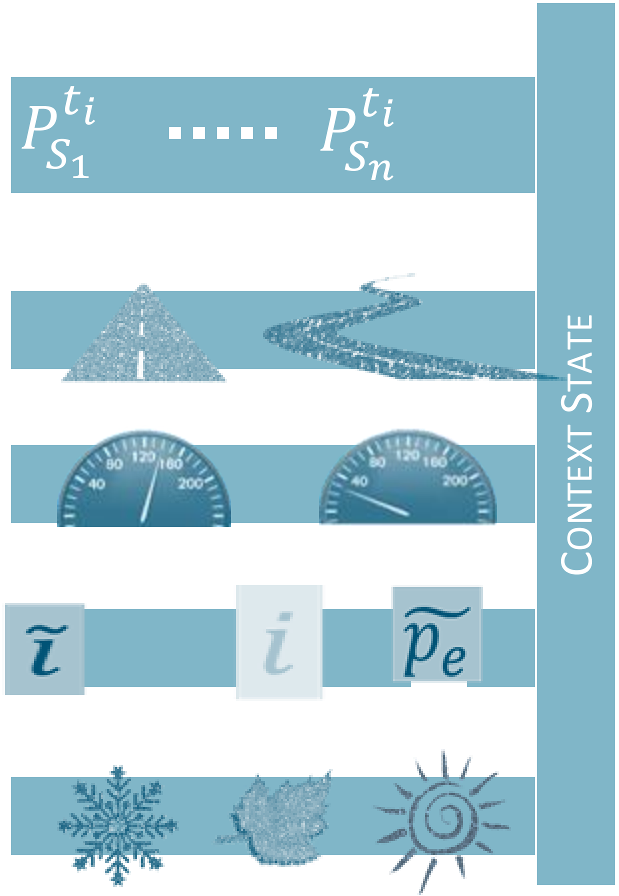
Function	Failure Mode	Situation	Exposure	Controllability	Severity	ASIL
ACC	Self-Acceleration	City, stopping at pedestrian crossing	E4	C3	S2	C
ACC	Self-Acceleration	Highway	E4	C1	S3	B

Safety Goal	ASIL
An unintended self-acceleration of more than 2 m/s^2 for more than 1 second must be avoided.	C

Shifting the model to runtime

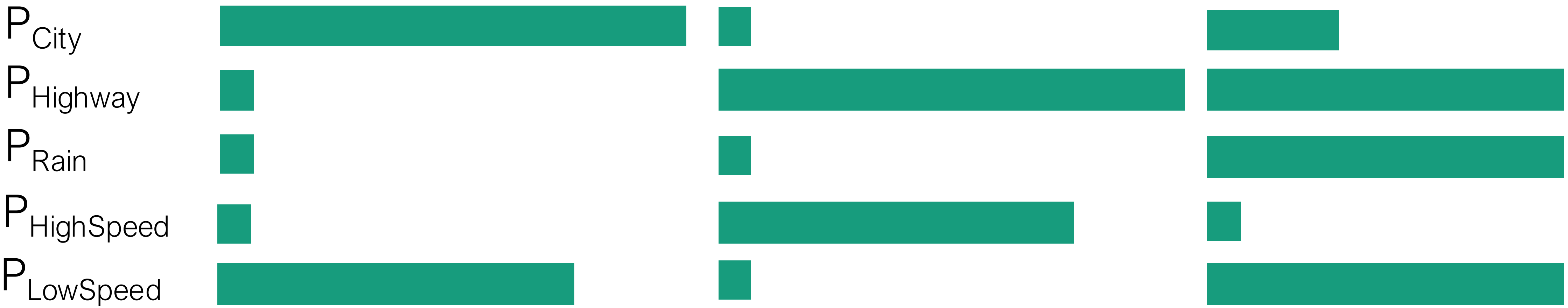


Hazard Analysis and Risk Assessment (HARA) @ Runtime

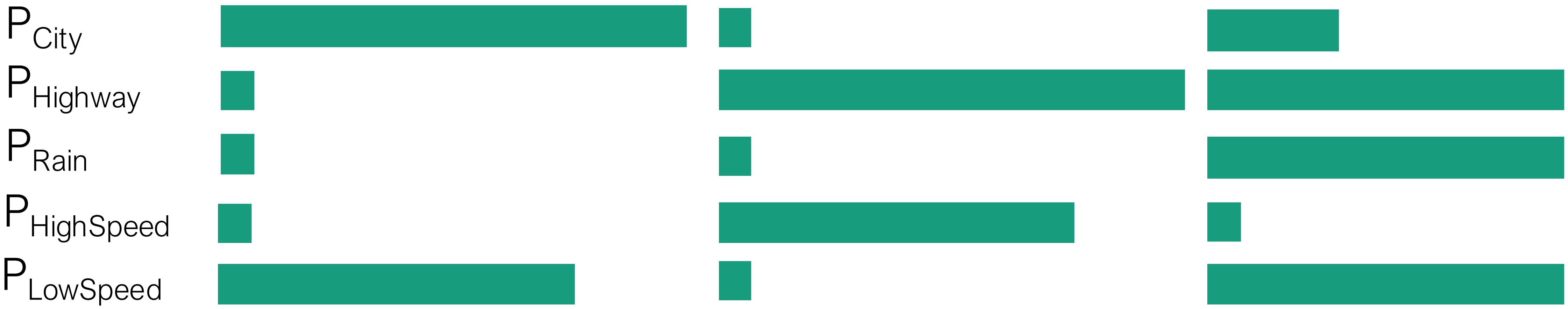


Exposure	Controllability	Severity	SafetyGoal	Integrity
$\widetilde{p_e}$	c	s	$\in \mid \notin$	i
$\widetilde{p_e}$	$\tilde{c}$	$\tilde{s}$	$\in \mid \notin$	$\tilde{\imath}$

The Example



The Example



Situation	Exposure	Controllability	Severity	ASIL
City, stopping at pedestrian crossing	E4 ✓	C3 ✓	S2 ✓	C ✓

C

Situation	Exposure	Controllability	Severity	ASIL
Highway	E4 ✓	C1 ✓	S3 ✓	B ✓

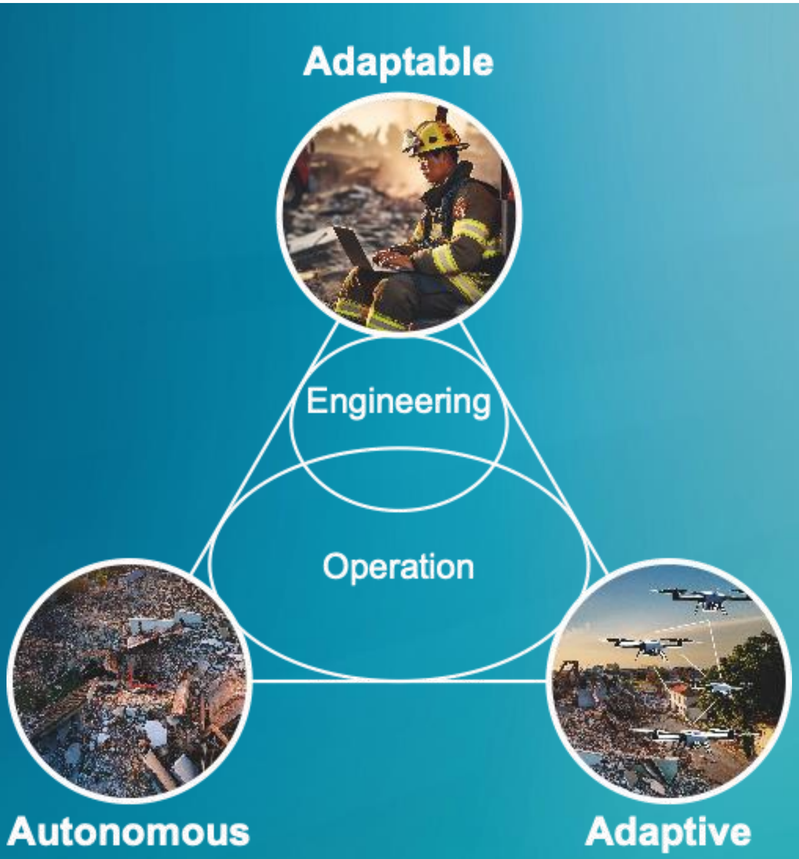
B

Situation	Exposure	Controllability	Severity	ASIL
Highway	E4 ✓	C2 ↘	S1 ↑	A ↗

A

Overview

An Overview on our Research



Overview

An Overview on our Research

Autonomous

Adaptive

Adaptive Intelligence

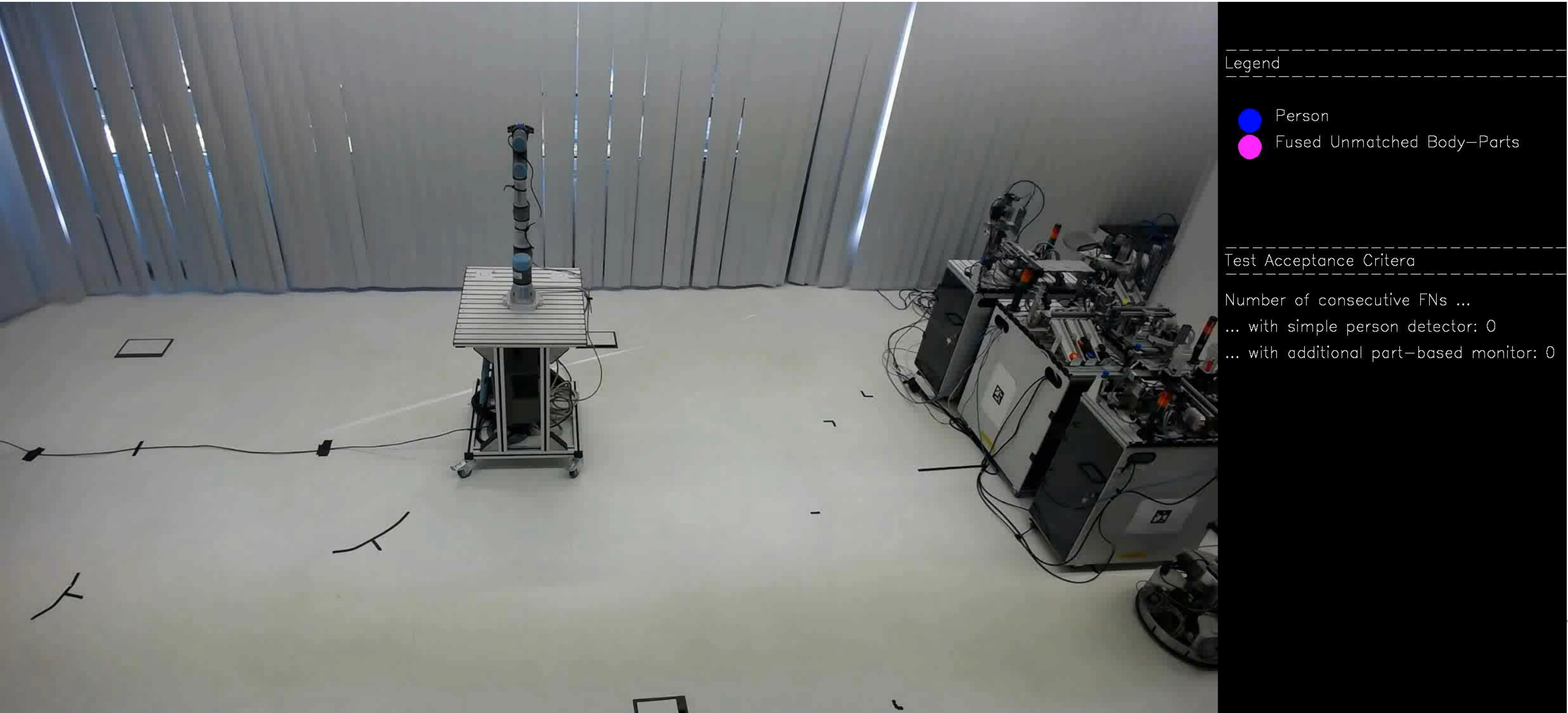
Dual Intelligence

Tell me the letters' color, not the word

Blue Red Green Blue

Red Green Blue Red

Dual Intelligence – Monitoring Architecture



Legend

● Person

● Fused Unmatched Body–Parts

Test Acceptance Critera

Number of consecutive FNs ...

... with simple person detector: 0

... with additional part–based monitor: 0

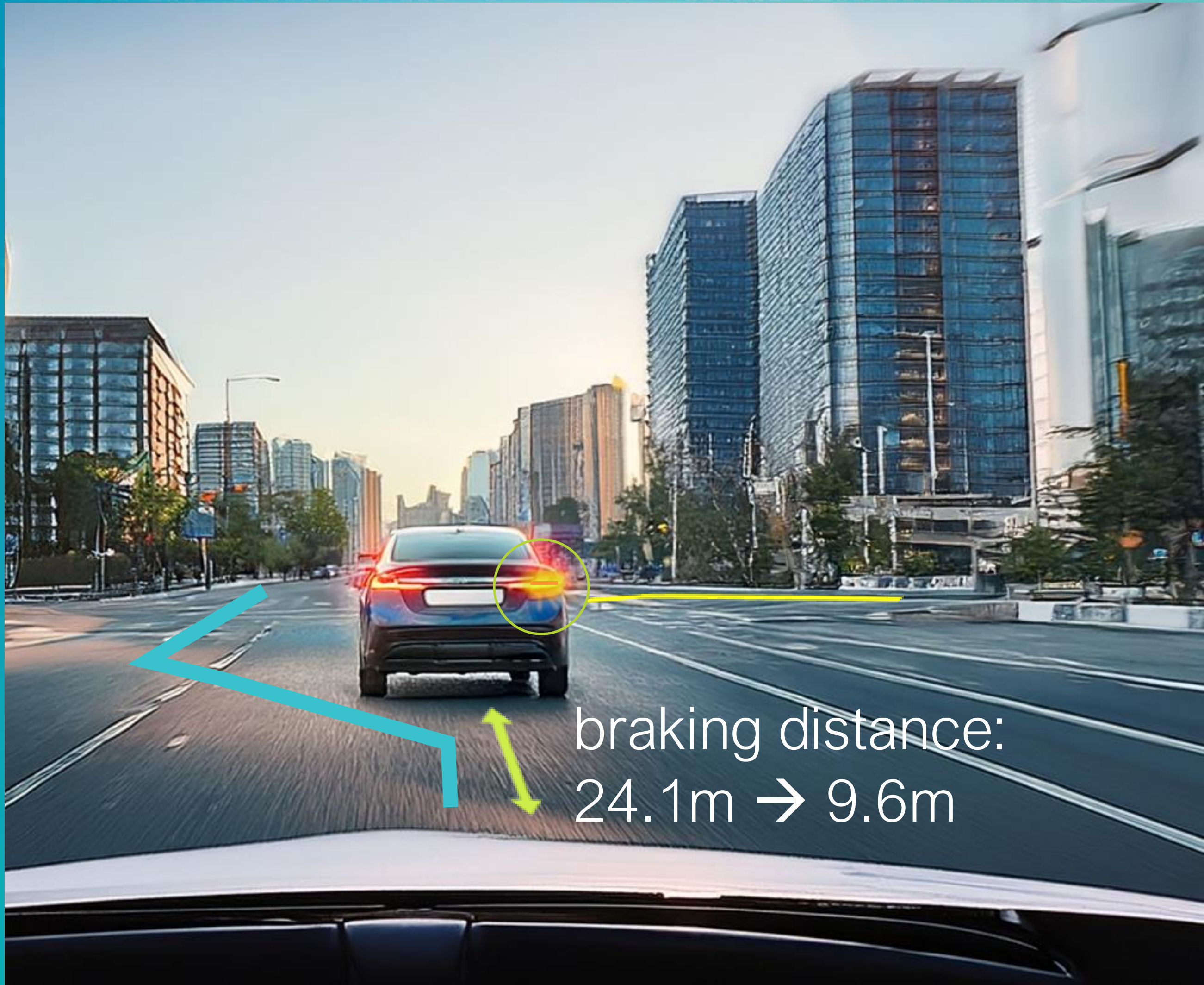
Intelligent Resilience

What a human driver would do

- Identification of traffic participants' intent
- Estimation of likelihoods and alternative options in case of misprediction
- Reassessment of risk
  → smooth driving over maintaining worst-case distance

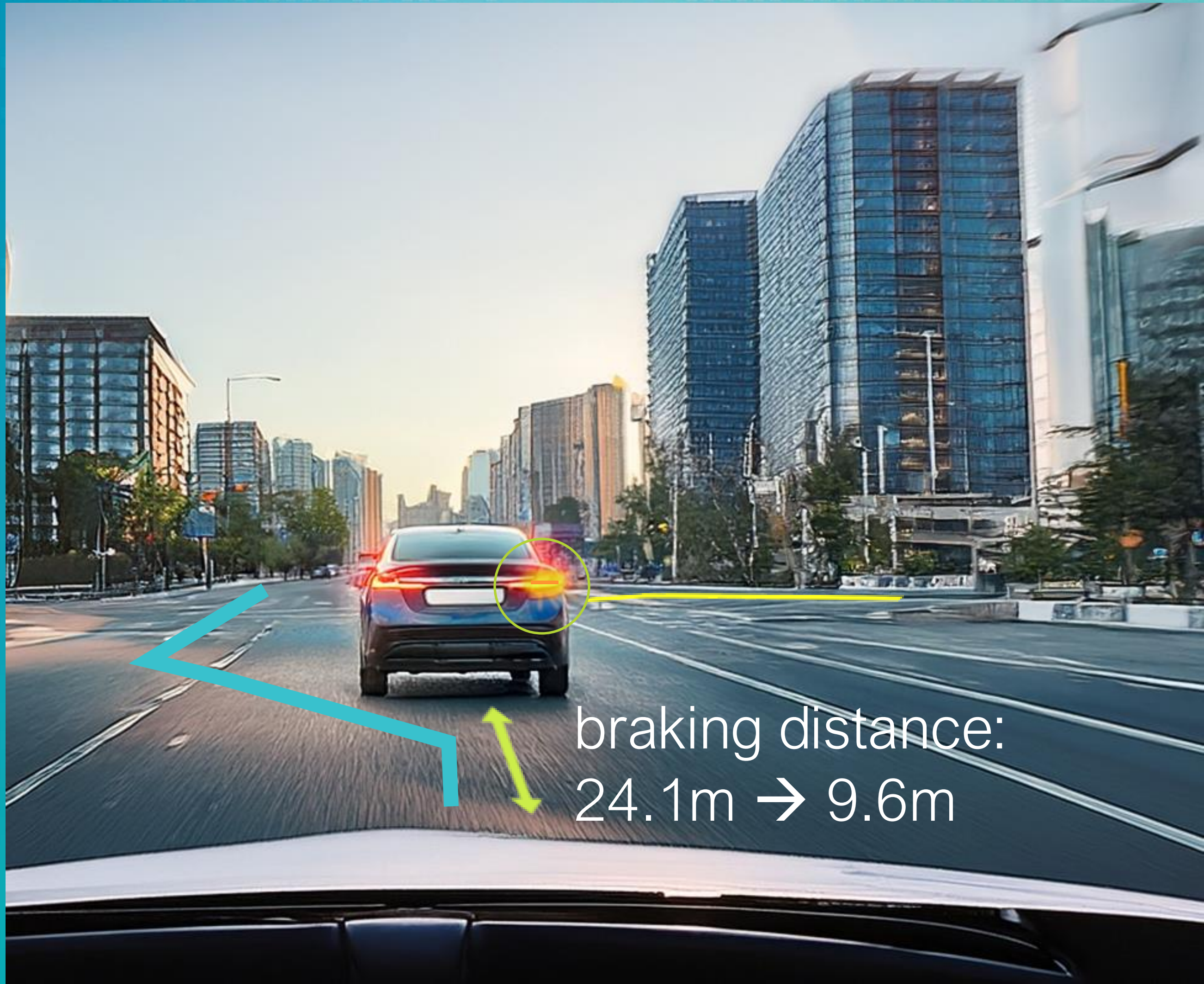
Intelligent Resilience

- Conservative safety monitors would not allow a violation of a worst-case minimum distance.
- We start mixing comfort and safety (e.g., what deceleration feels comfortable instead of what is physically possible.)
- But: Separation of concerns would give us additional freedom



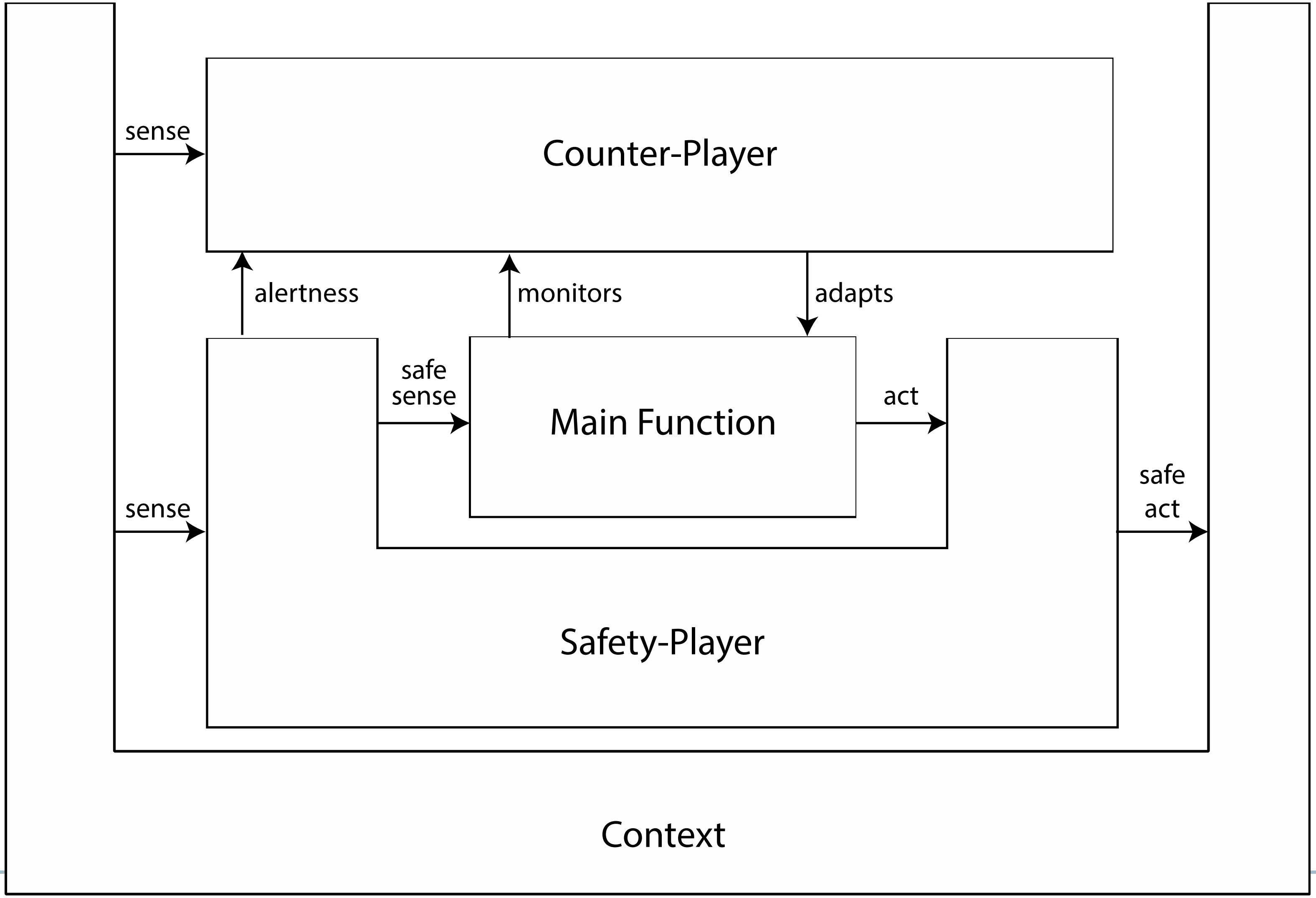
Intelligent Resilience

- Understanding the scene requires AI, V2X-data, cloud-data
- This technology could improve utility, but wouldn't be allowed in the safety-critical path
- How to exploit the potential of AI, cloud etc. without violating safety?

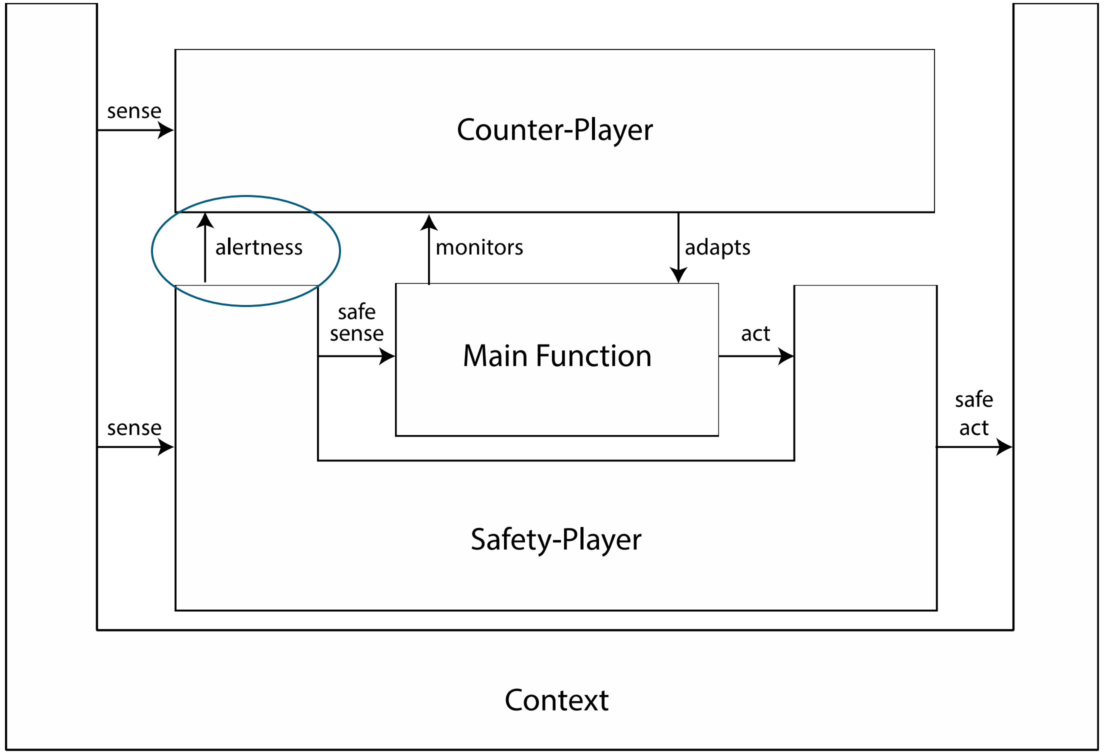


braking distance:
24.1m → 9.6m

The Safety-Counter-Player Architecture



The Safety-Counter-Player Architecture (cont.)



The counter-player "plays" against the safety-player by optimizing utility and minimizing the likelihood of an intervention of the safety-player.

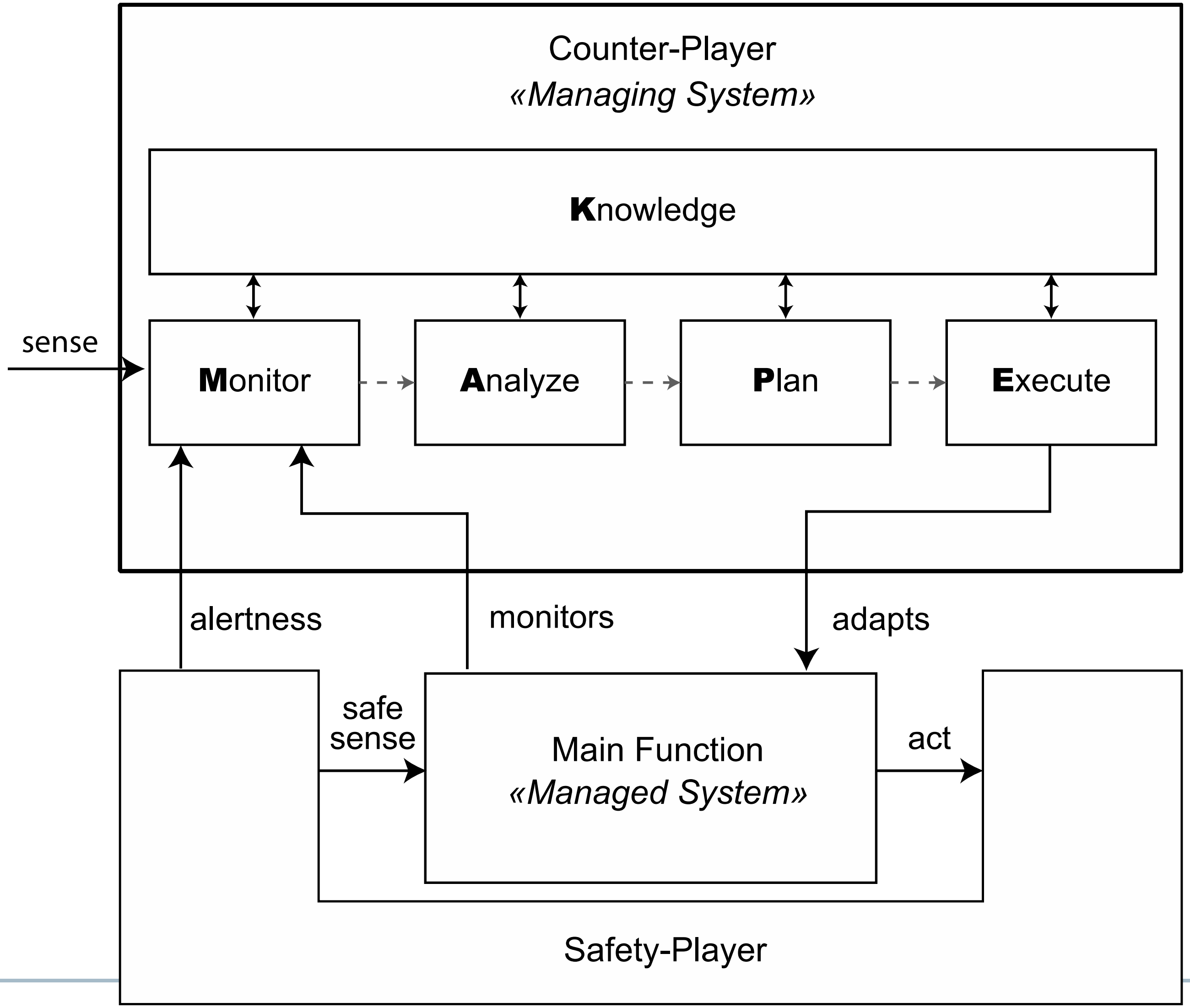
The counter-player is not within the safety-critical path.

The safety-player provides an alertness value [0,1] instead of a binary guard to allow the counter-player to adapt its strategy.

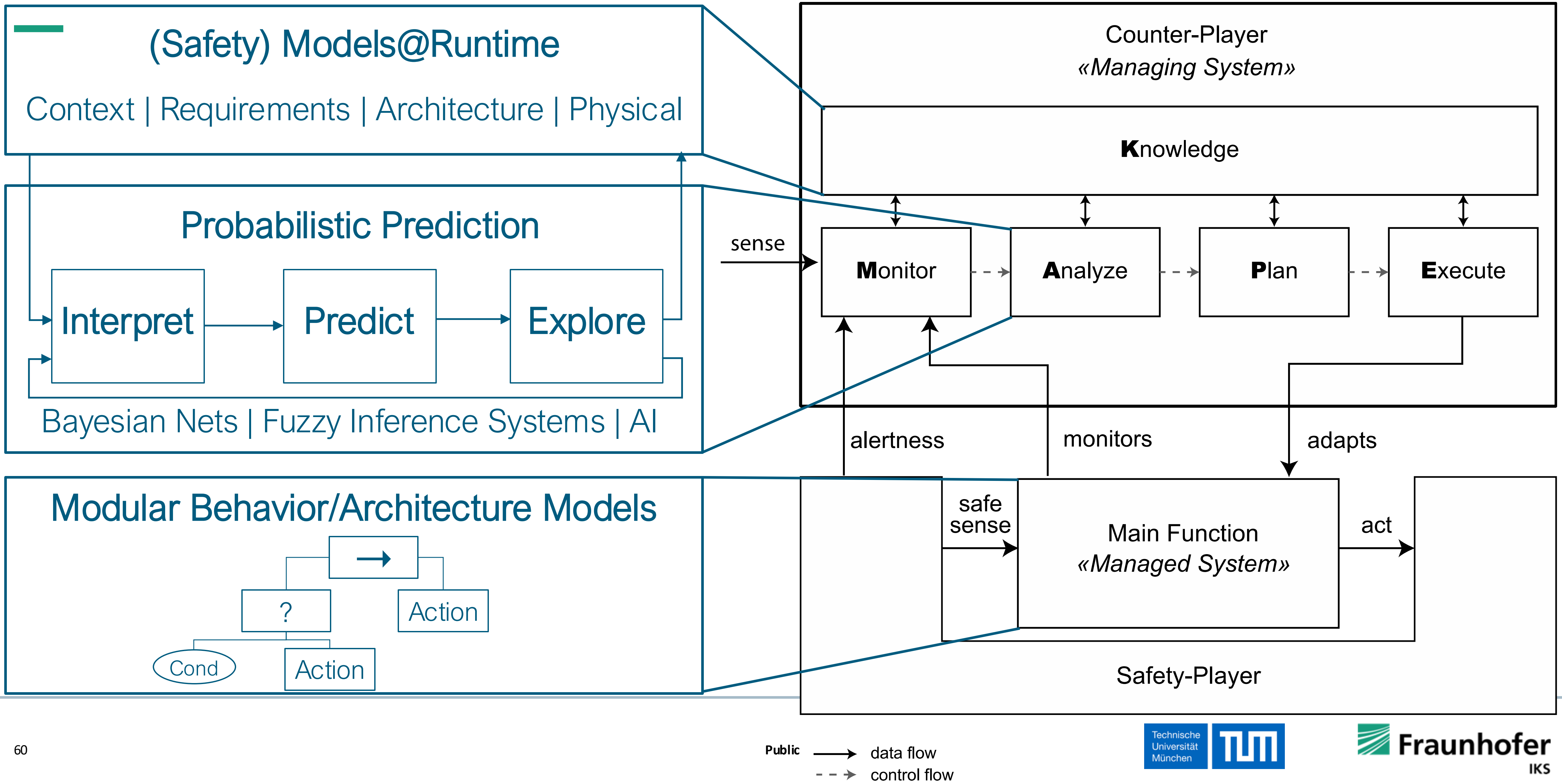
The safety player focuses on safety – and safety only – as "last line of defense".

Realization as Self-Adaptive System

- The counter-player still needs to be highly-reliable.
- It should follow basic principles of safety, going beyond what would be considered safe.
- High-Quality instead of "religious" safety.

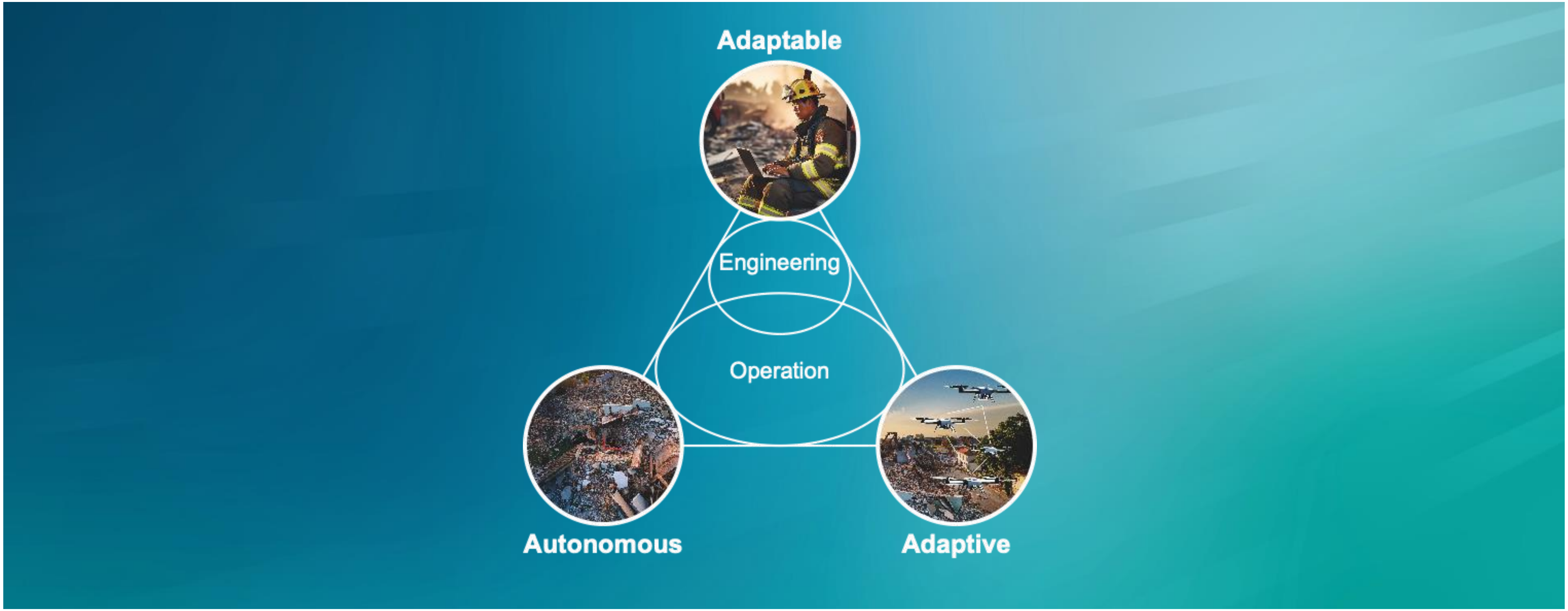


Dynamic Behavior Adaptation



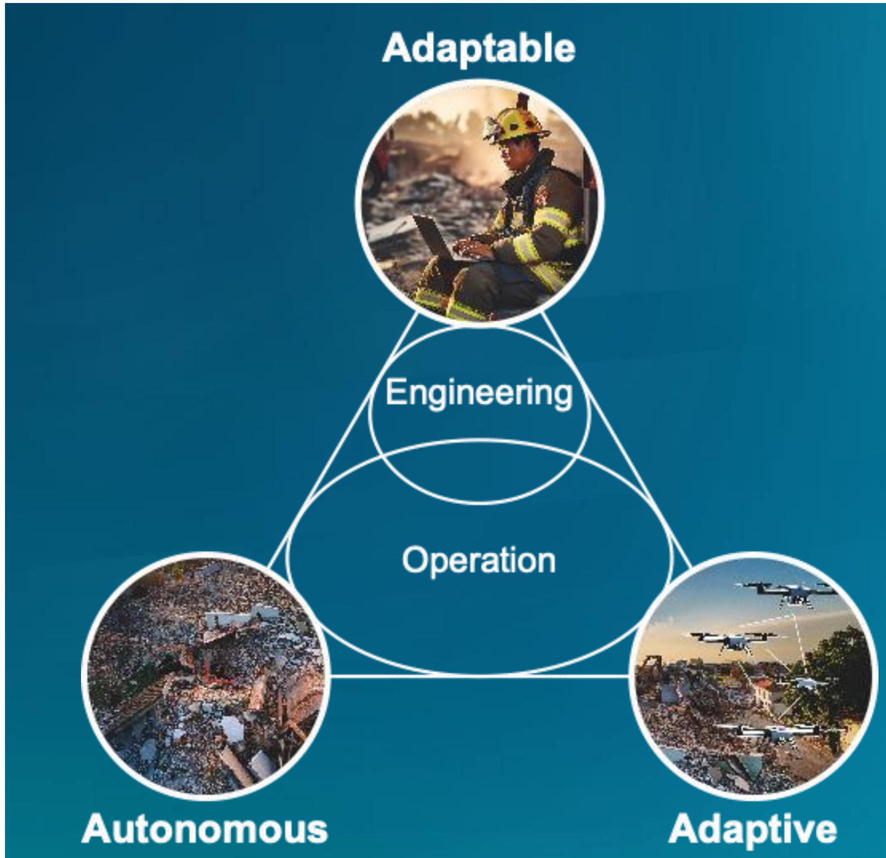
Overview

An Overview on our Research



Overview

An Overview on our Research



Triale Intelligenz

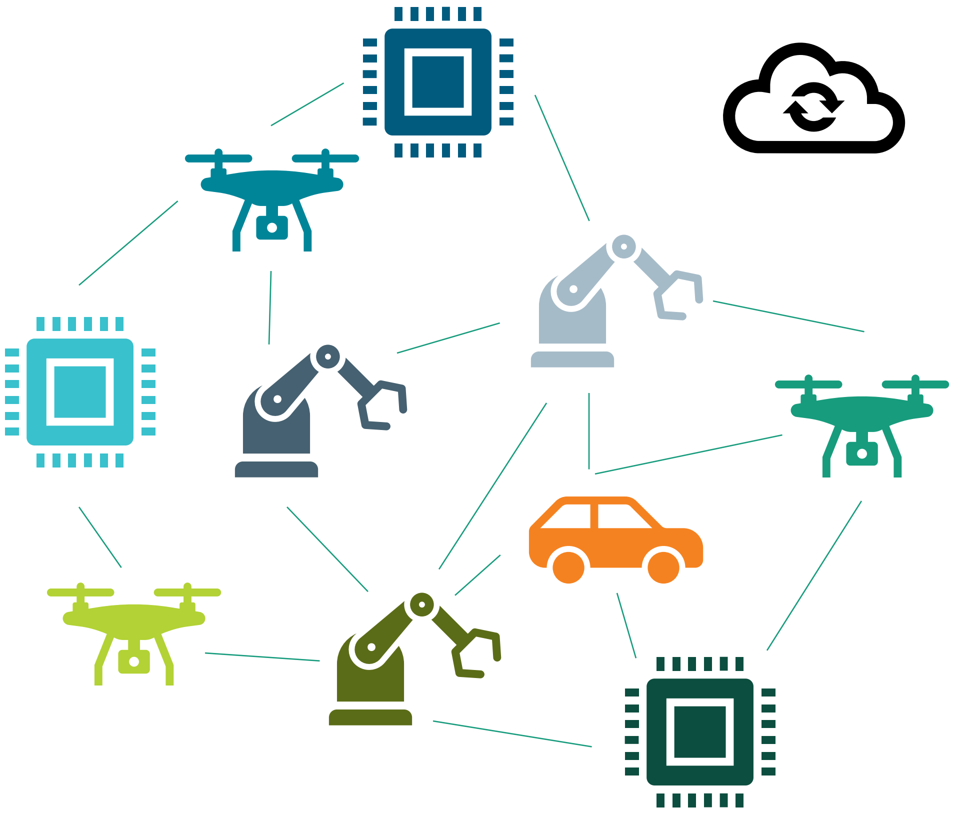
Three-Fold Intelligence

Understanding Human-AI-Collaboration

- Collaboration between AI and humans comes with many challenges. However, many of these challenges are estimated based on best guesses.
- Which methods of interaction reduce the human workload, and which ones increase it? Which types of interaction lead to complacency, and which do not?
- Surveys are often biased and influenced by psychological effects. For example, would you admit to merely clicking "approved" without truly reading the output?

➔ Our ongoing work aims to use brain-computer interfaces (BCI) to objectively measure real workload, attention, and other parameters. This analysis will help us identify the risks as well as the dos and don'ts of human-AI interaction.

Constitutional Safety Assurance



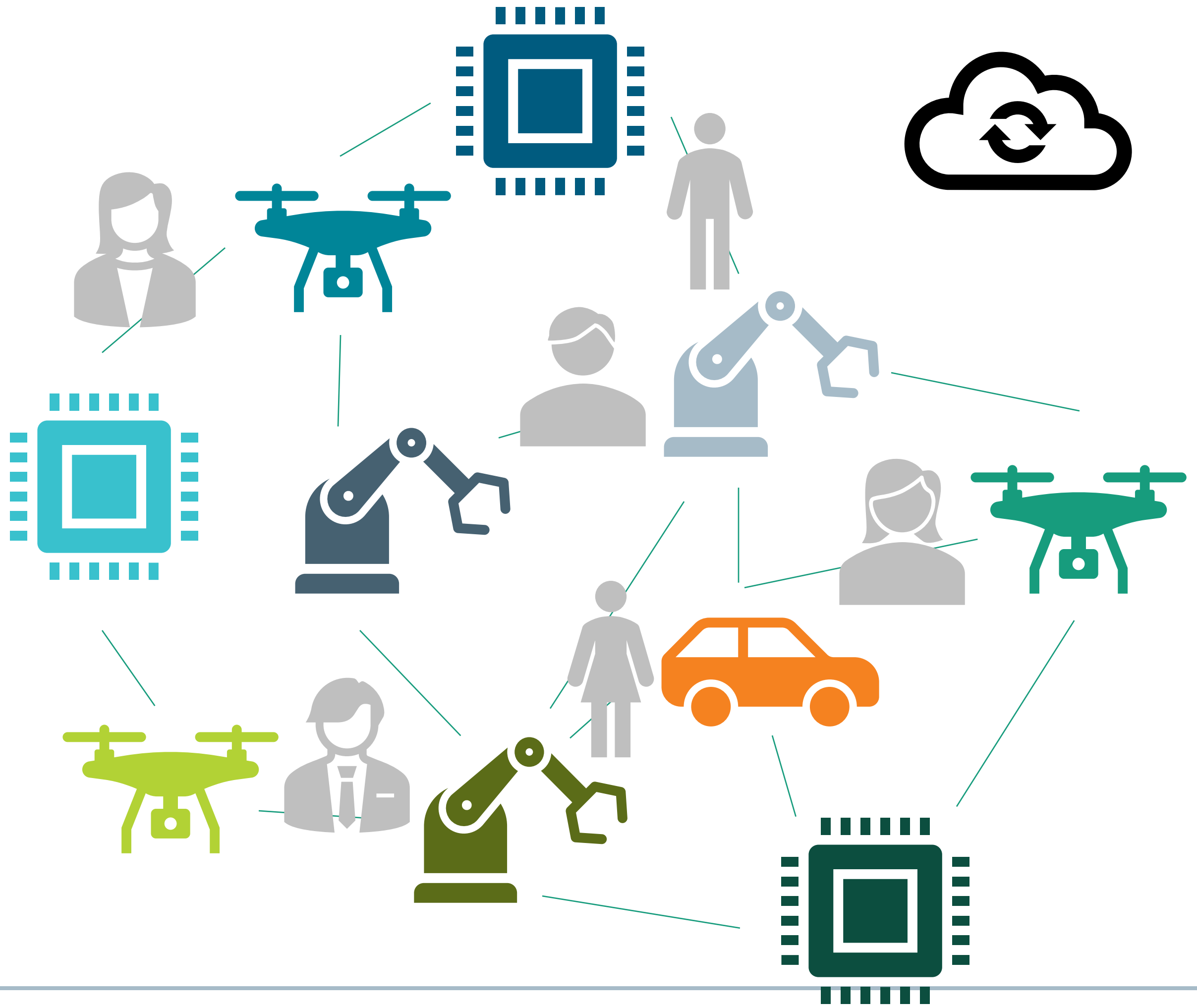
$$\lim_{n \to \infty} Complexity$$

Thought Experiment:

- Imagine a multitude of autonomous, independent agents,
- all interconnected with one another and the cloud.
- These agents must interact and collaborate to achieve a common goal.
- As the complexity of this system increases, it becomes impossible to maintain central control, as there is no single node responsible for ensuring the overall safety of the system.

Why do we think we can apply the same safety approaches used for a saw blade cover?

Constitutional Safety Assurance

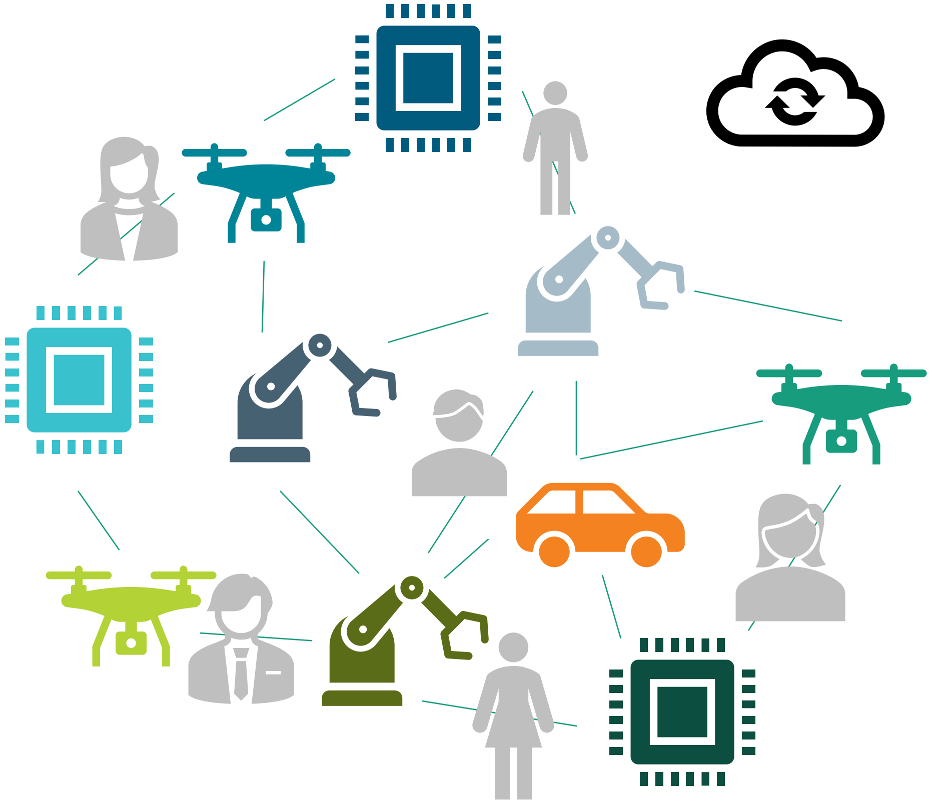


Furthermore imagine:

- A scenario where all systems are not only interacting but also collaborating with humans.
- The combined behaviors of both systems and humans result in the emergent behavior of the ecosystem, which drives the achievement of a common goal.
- The behavior of humans is influenced by the behavior of the systems, and (potentially) vice versa.
- This environment is rife with uncertainty and misunderstandings.

Doesn’t this resemble a team or a society of collaborating individuals?

Constitutional Safety Assurance

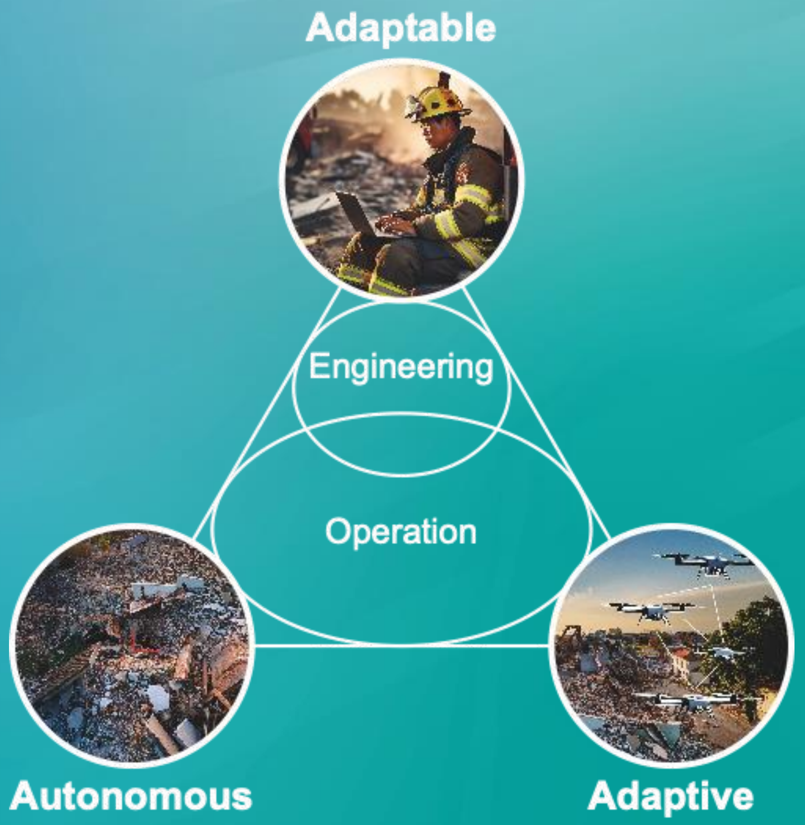


Wouldn’t it make sense to at least consider...

- Handling a complex system is akin to managing a society, utilizing principles that guide social interaction?
- To establish joint rules for “living” and “working” together – a **safety constitution**.
- Having a **legislative body**, such as operators and policymakers, that defines the constitution and laws.
- Implementing ecosystem components that function like an **executive** to ensure compliance with laws and intervene in cases of violations.
- A **judicial body** that punishes "criminals" by banning the systems / their manufacturers from the ecosystem.
- And numerous other analogies, such as rescue teams healing the ecosystem and preventing further damage…

Summary:

Let's focus on the big picture, not just the individual pieces.



Thank you

Prof. Dr. Mario Trapp

mario.trapp@iks.fraunhofer.de
mario.trapp@tum.de

Fraunhofer IKS

Fraunhofer Institute for Cognitive Systems IKS