safe.trAIn: Safety Assurance of a Driverless Regional Train

Keynote VEHITS/SMARTGREENS 2025 | 2nd April 2025 | Marc Zeller, Siemens AG

Funded by the European Union
NextGenerationEU
Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

safe·trAIn

Why safe.trAIn?

01

“Fully Automated Train Operation” is a significant lever to reduce CO_2-Emissions because it supports the shift from individual to public transport

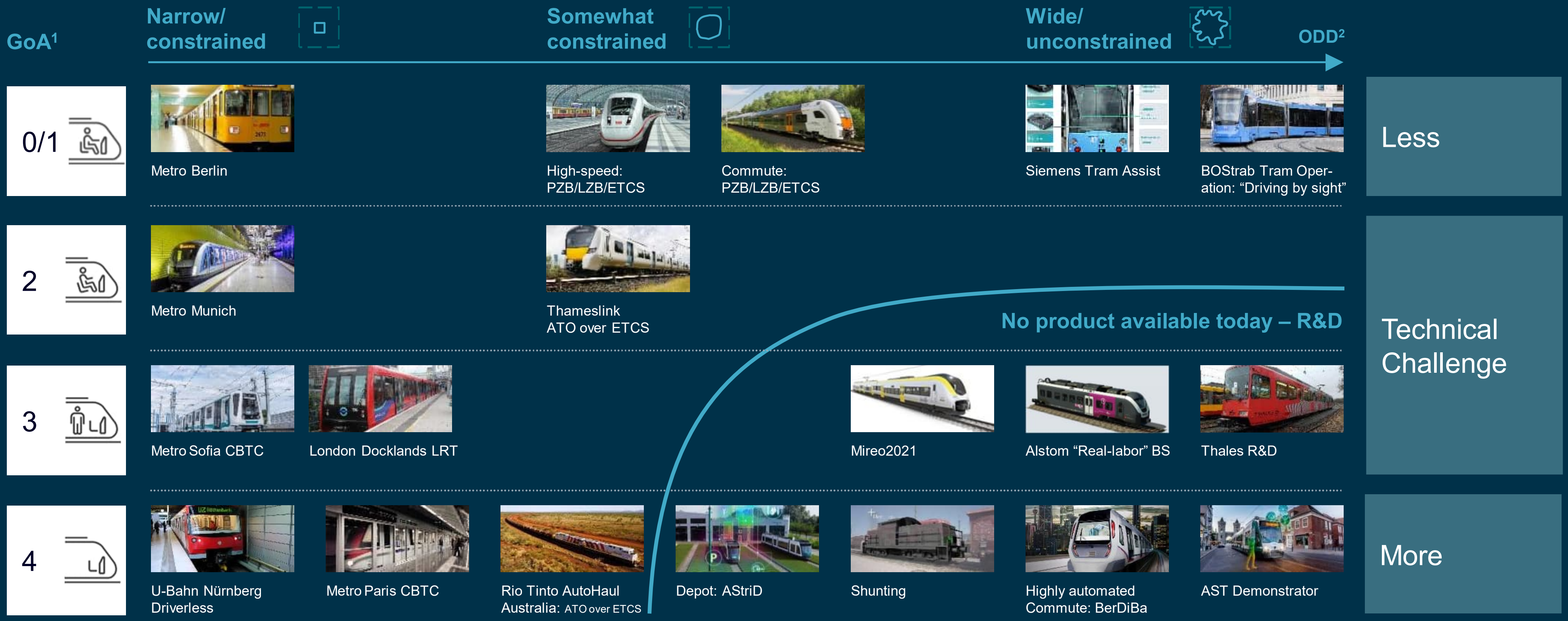
- Overcoming the considerable shortage of train drivers
- Reduction of unproductive times (paths from the train driver (TF) to and from the vehicle)
- Densification of the timetable, e.g., by splitting vehicles that would other-wise run in multiple traction or additional connections in off-peak times
- Increased flexibility in timetable design
- Faster achievement of normal operation in the event of malfunctions, as replacement vehicles are provided more rapidly

Steps in the introduction of highly and fully automated driving

Manual operation Supervision by driver	**Highly automatic operation** Limited driver action	**Fully automatic operation** No supervision by driver	
GoA 1 Manual train operation with driver Supervision and control train operation (SCO)	**GoA 2** Automatic train operation with driver Semi-automated train operation (STO)	**GoA 3** Automatic train operation without driver Driverless train operation (DTO)	**GoA 4** Automatic train operation without staff Unattended train operation (UTO)
Provision of driving recommendations for energy-optimized train runs			
Driver drives completely manually	Automatic train operation after driver interaction	Automatic train operation	
Obstruction detection by driver	●	Automatic obstruction detection (obstacle detection, platform protection)	
Manual train dispatching by driver or train attendant	●	●	Central or automatic train dispatching
Train monitoring and intervention in emergency situations by driver or train attendant	●	●	Central monitoring or automation functions for handling of train disturbances and emergency situations

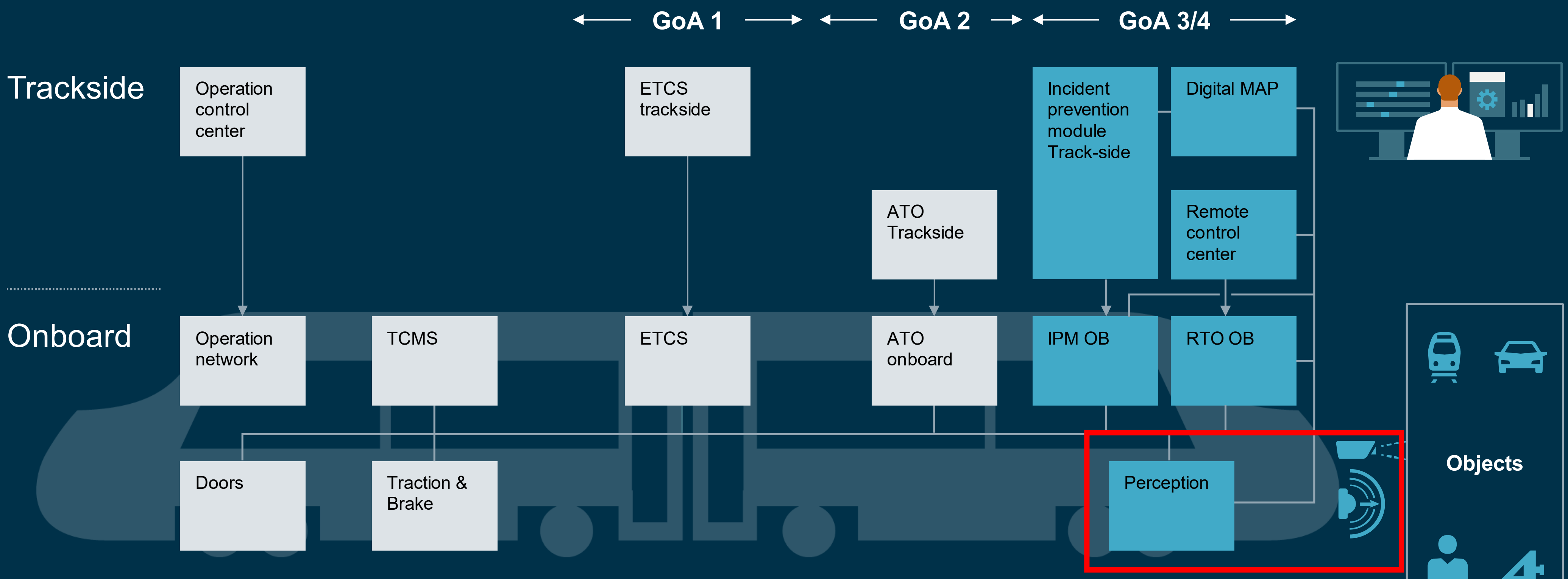
GoA Grade of Automation acc. to IEC 62267

Automation in the Railway Domain



1 GoA = Grade of Automation (acc. to IEC 62290) | **2** ODD = Operational Design Domain = Operation conditions under which an autonomous system is specifically designed to function

Establishment of a system for GoA 3/4 operation



Technical key challenge: Obstacle/object detection in a natural environment

Up to GoA 2 components
Added components for GoA 3/4

Artificial Intelligence (AI) is required to enable object classification in open environments, but it is a big challenge to build dependable systems

According to the current state of the art, we assume that e.g., obstacle detection can only be implemented using the methods of ML (Machine Learning – a field of AI Artificial Intelligence).

Legislation

Safety Standards

Processes und tools

New: European AI ACT- high-risk application

No safety standard for AI-based perception system in rail domain

Unclear regarding inter- pretability, robustness

No established tools and processes

AI is a "black box"

Dynamic open environ- ment

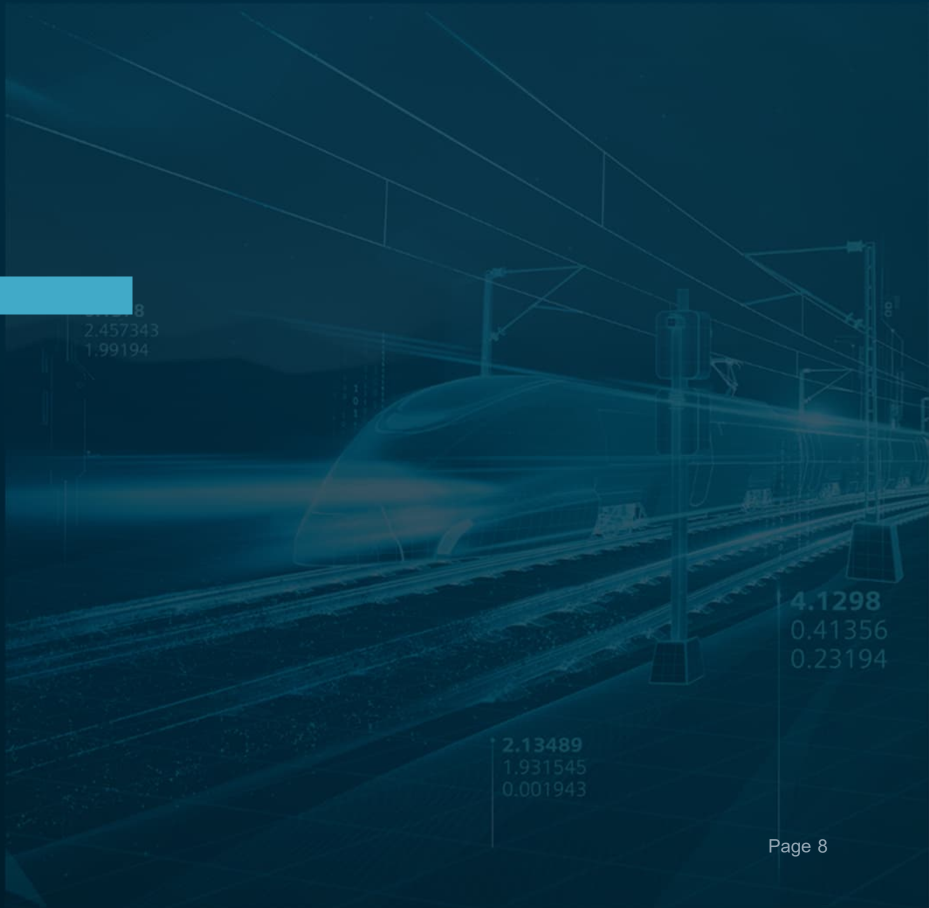
Requirements for assessment

Interpretability

Uncertainty

What is safe.trAIn?

02



Safe.trAIn enables Safe Perception for Driverless Regional Trains

Challenges of AI in Railway

- No safety standard for AI-based perception in rail domain
- Unclear requirements for assessment of AI
- No established tools and processes

Project goals

Safe perception for automated trains

Safety-enabling architecture
Exploration of architecture patterns involving redundancy

Metrics/KPIs for (self)-evaluation
Performance metrics for online and offline evaluation

Safety case and testing
Quantitative evaluation of all approaches in virtual test field

Transfer to standardization
Contributions to national and European standardization activities

Consortium

Technology Provider

Assessor/ Standardization

User

Enabler

Funded by the European Union
NextGenerationEU

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

Person on track and passenger in train are the 2 safety objectives for perception system

Passenger in train

The perception system will detect heavy obstacles on the tracks, a collision with which can potentially cause injuries and fatalities for passengers in the train

Heavy obstacles include, but are not limited to trees, rocks, cars, trucks, other trains, flooding, landslide…

Current safety objective of the rail operation acc. to German regulations (e. g. DB RIL 408.2341) The driver must prevent harm from the train.

Safety objectives

The perception system will prevent harm from passengers in the vehicle and persons on the track

Person on track

The perception system will detect persons on the tracks, a collision with which can potentially cause injuries and fatalities for the **person on the track**

Persons on the track include, but are not limited to workers, trespassers, playing kids, …

Probably needed for public acceptance of driverless train operation.

It is challenging to match safety requirements with AI-related evidences

Safety Requirements for a specific application
(Safety Functions with Safety Integrity Level)

Independent of technology,
i.e., whether AI is used or not

How does that match?
To be demonstrated for the specific case, no generally accepted “recipe” for AI fulfilling SIL exists in standards

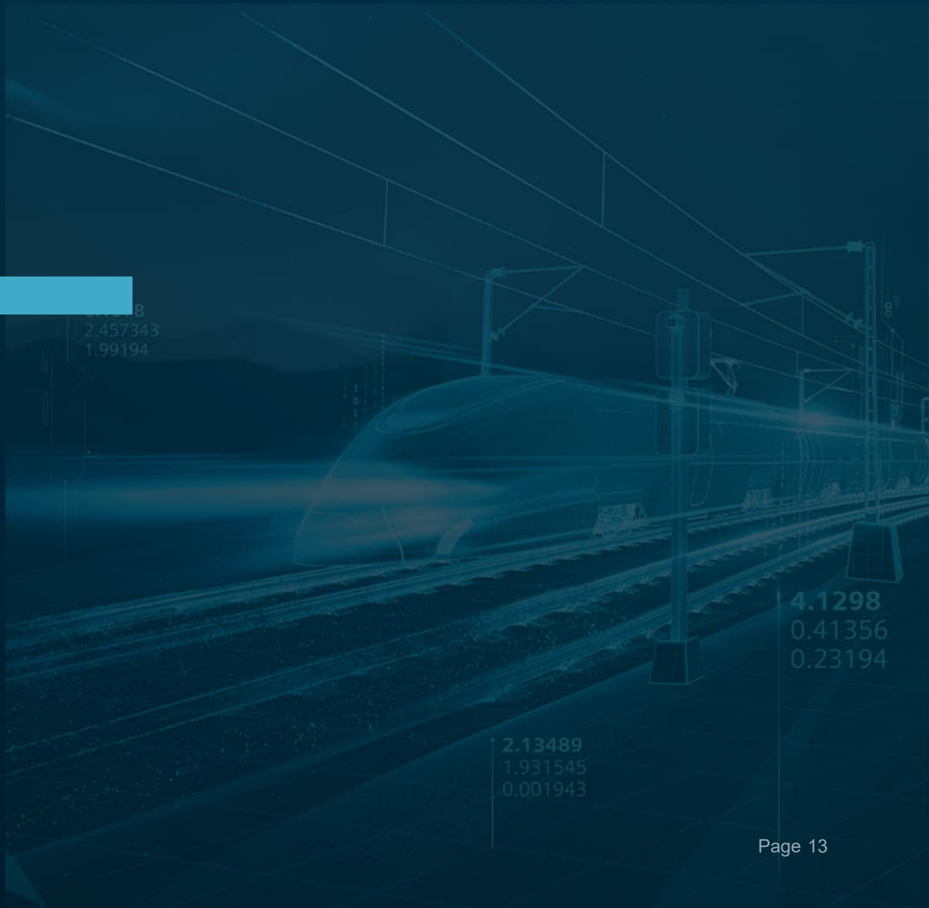
Evidence from
Machine Learning specific properties, metrics, thresholds, …

Is this really “evidence”?
For what?

ISO/IEC TR 29119-11:2020 Guideline on the testing of AI-based systems:
“The currently available AI frameworks and algorithms are **not qualified** for use on the development of safety-related systems.”

What did we achieve in safe.trAIn?

03



The overall safety target relates to the concept of Recall

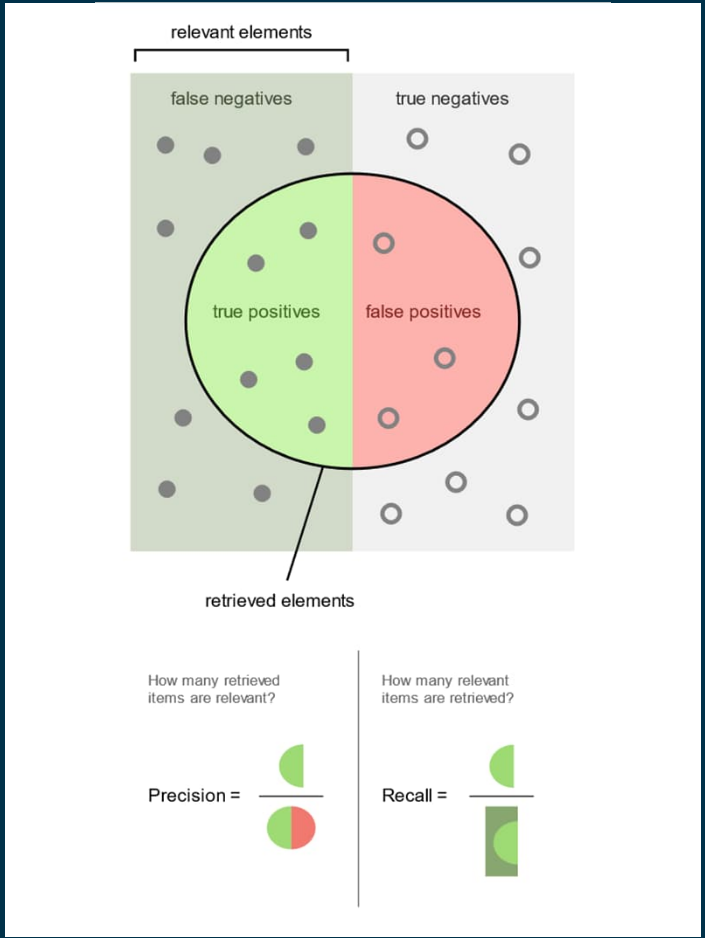
According to CSM RA “comparison with reference system”

- **Safety target: “overall as good as driver”**

  Regional trains rarely encounter Obstacles

  → Evaluate safety against Probability of Failure on Demand (PFD)

- **PFD = 1%**
  - Based on ATO-Risk[1] project and further analysis
  - PFD is considered as equivalent to 1–recall, where recall =TP/(TP+FN)
  - TP and FN to be evaluated against definition of safety functions
  - Achieved PFD will be determined offline using validation data with ground truth
  - Recall to be evaluated on set of scenarios



Source: Wikipedia

[1] https://www.dzsf.bund.de/SharedDocs/Downloads/DZSF/Veroeffentlichungen/Forschungsberichte/2023/ForBe_40_2023_ATO_Risk_Summary_EN.pdf?__blob=publicationFile&v=5

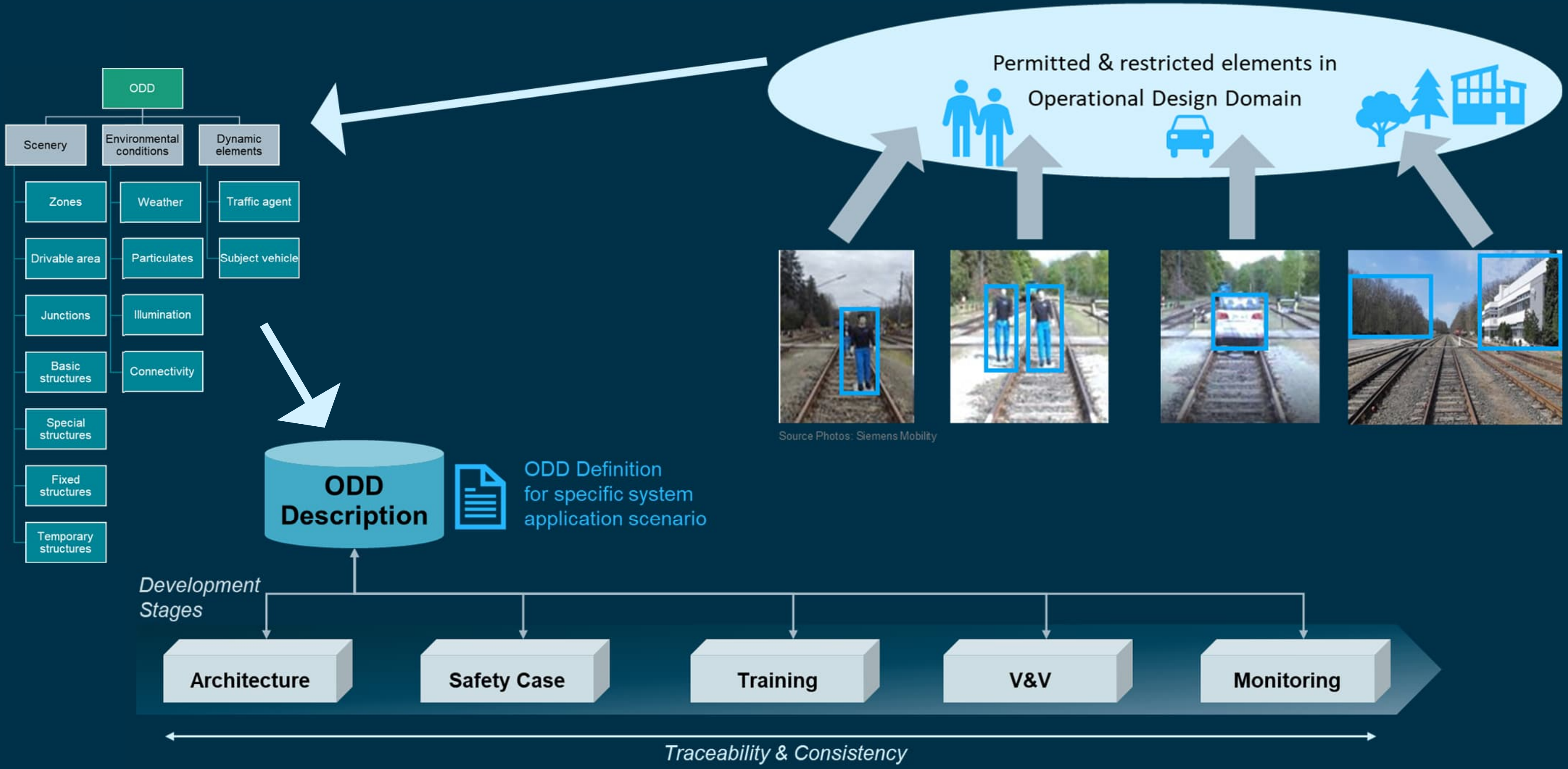
Five Pillars of Safety Case Strategy address different aspects and must be balanced for specific circumstances

Safety Case Strategy

Processes tailored to applied perception specifics	Analysis of non-conventional redundancies in safety architecture	Demonstrate sufficient understanding of causalities of functional behavior	Testing with real and simulated data (in our virtual test field)	Safety Monitoring during Operation (e.g., Out-of Distribution Detection, …)
The defined processes needs to cover all developmental aspects considered important for the final assurance of correct behavior of the system under consideration.	In order to achieve a low-enough false negative rate, PFD, the architecture of the system comprises redundancies to cope with faults and imperfection of different perception components.	Its goal is to demonstrate a certain level of human understanding as to why the **right** results are given by the system for the **right** reasons.	Besides process reviews, audits, checking all documents, Q-Gates, etc. tests according to an acceptable coverage criterion are required.	Higher uncertainty in functional decision and behavior and possible domain shifts needs to be compensated by more stringent field monitoring compared to conventional system.

System Definition and Requirements

Operational Design Domain (ODD) as Central Element in the Development Process



Weiss G., Zeller, M., Schoenhaar H., et al. *Approach for Argumenting Safety on Basis of an Operational Design Domain*. In: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN ,24), 184–193 (2024). https://doi.org/10.1145/3644815.3644944

Pillar 1: To close the gap between assuring AI-based systems and conventional software systems: All AI Safety Concerns need to be addressed

Definition of AI Safety Concerns: **"AI-specific, underlying issues that may negatively impact the safety of a system."**
The AI Safety community has conducted comprehensive research on identifying AI Safety Concerns[1,2,3]:

AI Safety Concerns[1]

Inadequate specification of ODD	Inadequate planning of performance requirements	Insufficient AI development documentation	Inappropriate degree of transparency to stakeholders	AI-related hardware issues	Choice of untrustworthy data source	Missing data understanding
Discriminative data bias	Inaccurate data labels	Insufficient data representation	Inappropriate data splitting	Problems with synthetic data (Reality Gap)	Poor model design choices	Over- and underfitting

Lack of explainability	Unreliability in corner cases	Lack of robustness	Uncertainty concerns (model)	Integration issues	Operational data issues	Data drift (over time)	Concept drift

1 Schnitzer, R., Hapfelmeier, A., Gaube, S., Zillner, S.: AI Hazard Management: A framework for the systematic management of root causes for AI risks. | **2** Houben, S., Abrecht, S., Akila, M., Bär, A., Brockherde, F., Feifel, P., et al.: Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety. | **3** Willers, O., Sudholt, S., Raafatnia, S., Abrecht, S.: Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in SafetyCritical Perception Tasks

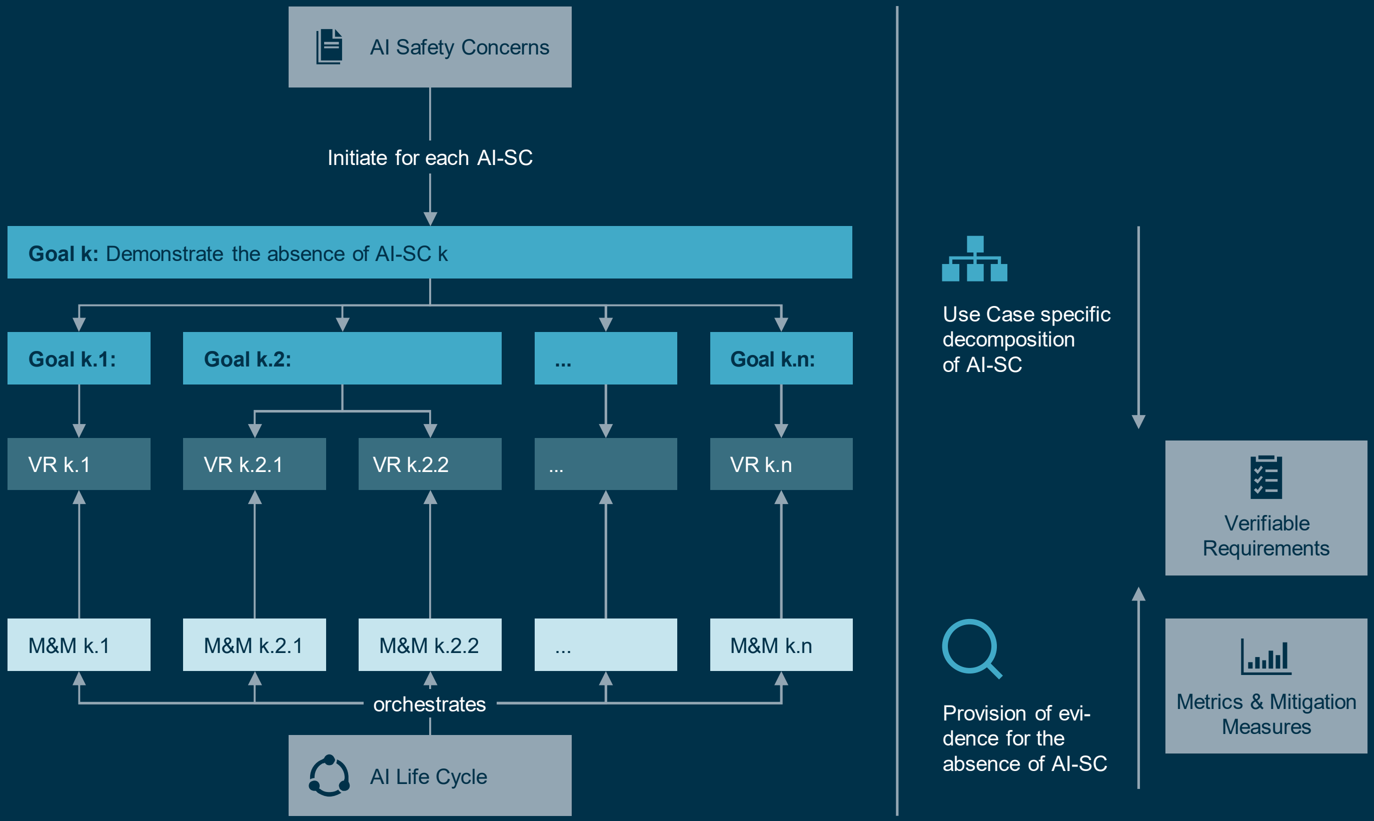
Pillar 1: Applying the Landscape of AI Safety Concerns consists of four steps

01 Initializing LAISC
Identification of relevant AI Safety Concerns

02 Decomposing the AI-SC
Use-case specific concretization of AI Safety Concerns

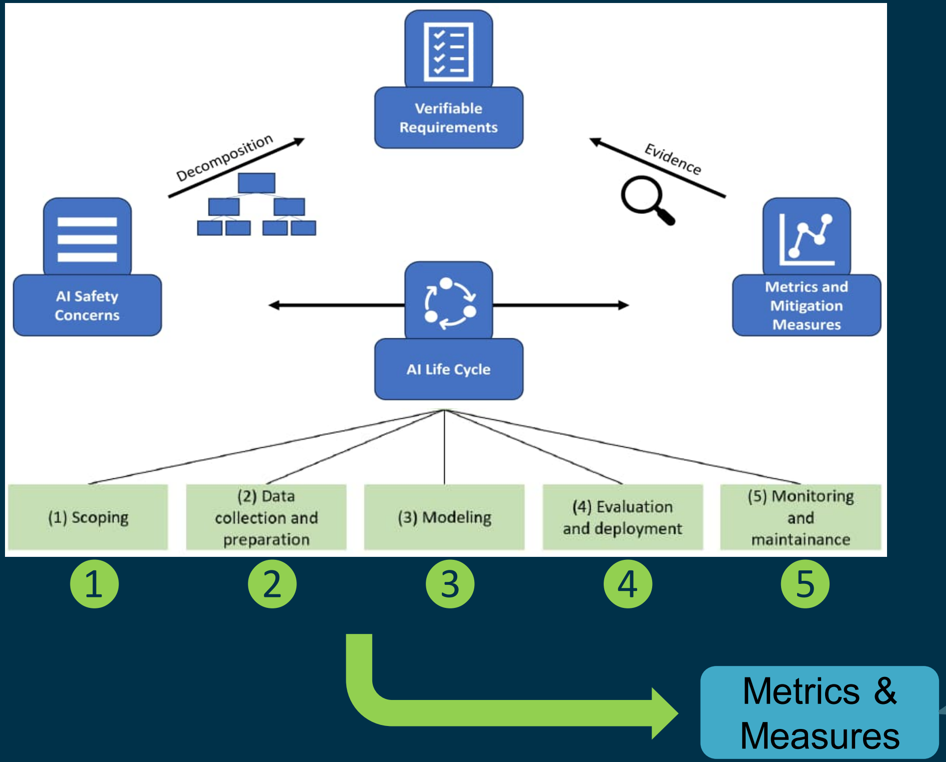
03 Derivation of Verifiable Requirements
Establishing criteria for determining when AI Safety concerns are considered absent

04 Application of Metrics and Mitigation Measures along the AI life cycle
Generation of evidences along the whole AI life cycle



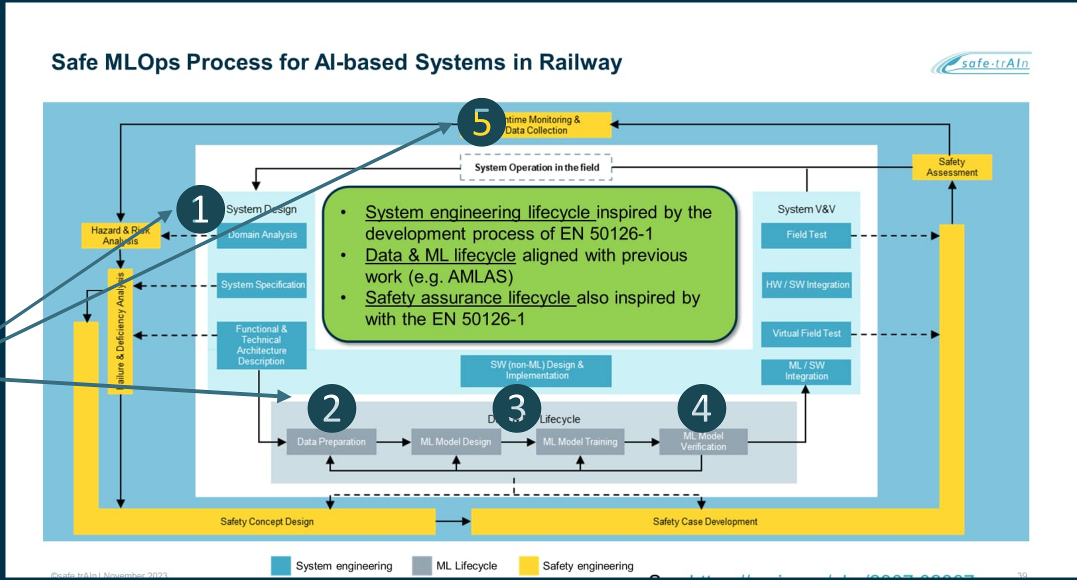
More details: Schnitzer, R., Kilian, L., Roessner, S., Theodorou, K., & Zillner, S. (2024). Landscape of AI safety concerns-A methodology to support safety assurance for AI-based autonomous systems. 8th International Conference on System Reliability and Safety (ICSRS) preprint available: https://arxiv.org/abs/2412.14020

Pillar 1: Landscape of AI Safety Concerns and safe MLOps Process



In order to assure AI-based autonomous systems:

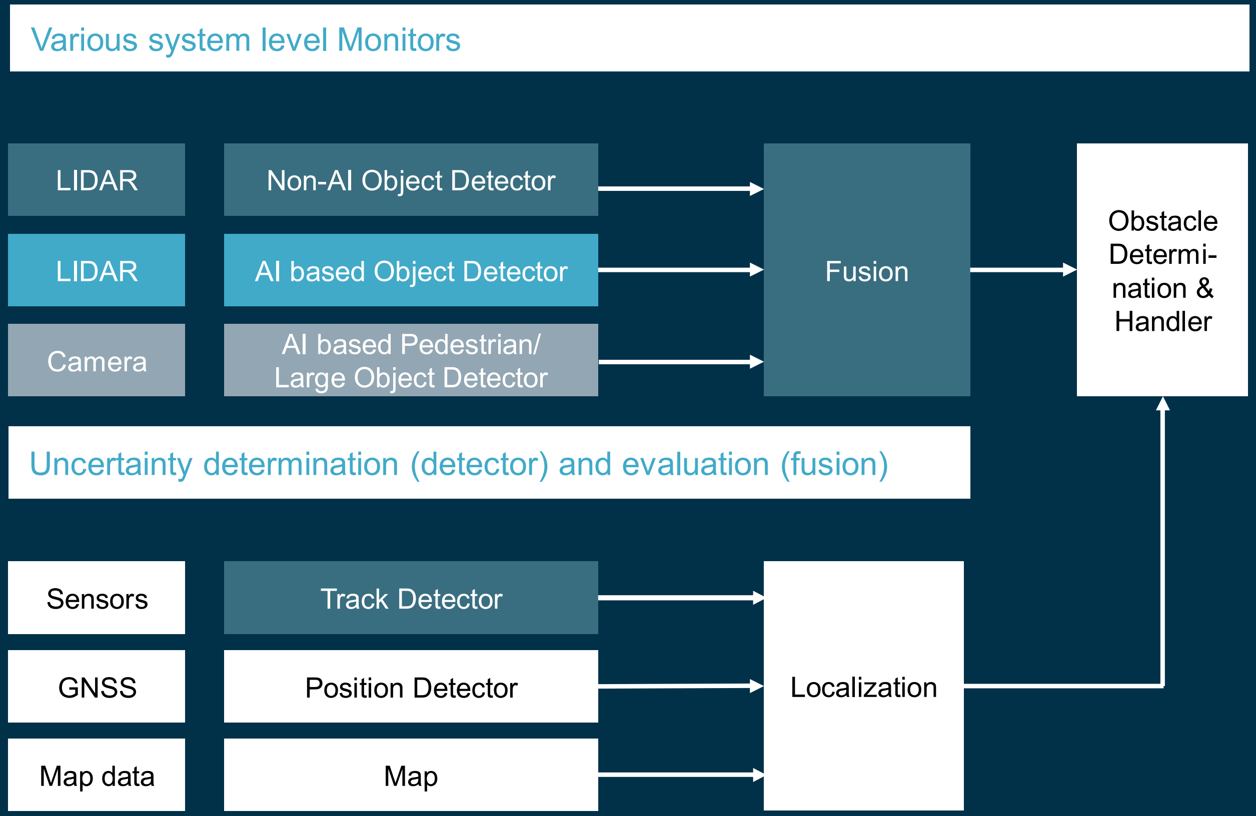
For each **AI Safety Concern, evidence** needs to be derived along **the whole AI life cycle** that **convincingly demonstrates** the sufficient mitigation of the respective AI Safety Concern.



More details: Schnitzer, R., Kilian, L., Roessner, S., Theodorou, K., & Zillner, S. (2024). Landscape of AI safety concerns-A methodology to support safety assurance for AI-based autonomous systems.
8th International Conference on System Reliability and Safety (ICSRS) preprint available: https://arxiv.org/abs/2412.14020

Zeller, M., Waschulzik, T., Schmid, R. et al. *Toward a safe MLOps process for the continuous development and safety assurance of ML-based systems in the railway domain.* AI Ethics 4, 123–130 (2024). https://doi.org/10.1007/s43681-023-00392-4

Non-conventional redundancies and *Monitoring* from Pillar 2 + Pillar 5



Define dissimilar architecture elements and data paths using

- Different sensor modalities
- Different detectors using AI and non-AI algorithms

Uncertainty determination and propagation partially implemented, e.g., by High Level fusion

Monitoring of system and components at runtime

- Safety measures realized in monitors and components

Pillar 3: Sufficient Understanding of Causalities of Functional Behavior is achieved by collaboration of AI and domain experts

Approach

- **Goal:** provide transparency and trust in the system's decision-making by demonstrating sufficient understanding of the causalities behind the perception system's functional behavior
  “**Does it do the right things for the right reasons?**”
- Focuses on **analyzing why the perception system makes certain decisions**, rather than just which decisions it makes. This includes – as far as possible – identifying potential biases or confounding factors
- **Limitation:** While full end-to-end explainability is not feasible, this pillar calls for providing appropriate levels of observability and explainability at the component level, using techniques like TCAV, Layer-wise Relevance Propagation (LRP) and Saliency Maps

Process

1. For each component, observability at the input and output interfaces proportionate to its influence on safety must be implemented
2. For each component, appropriate methods for explainability or interpretability are implemented, if possible and meaningful
3. Detailed behavior validation by a domain expert, supported by a perception system expert, must show evidence of the system’s suitability for use

This pillar focuses on leveraging both domain *and* perception system experts to review the system’s behavior comprehensively, ensure – as much as possible – that the perception system does the right things for the right reason.

Pillar 3: Saliency Maps help identify importance of regions of interests for the prediction

Objective

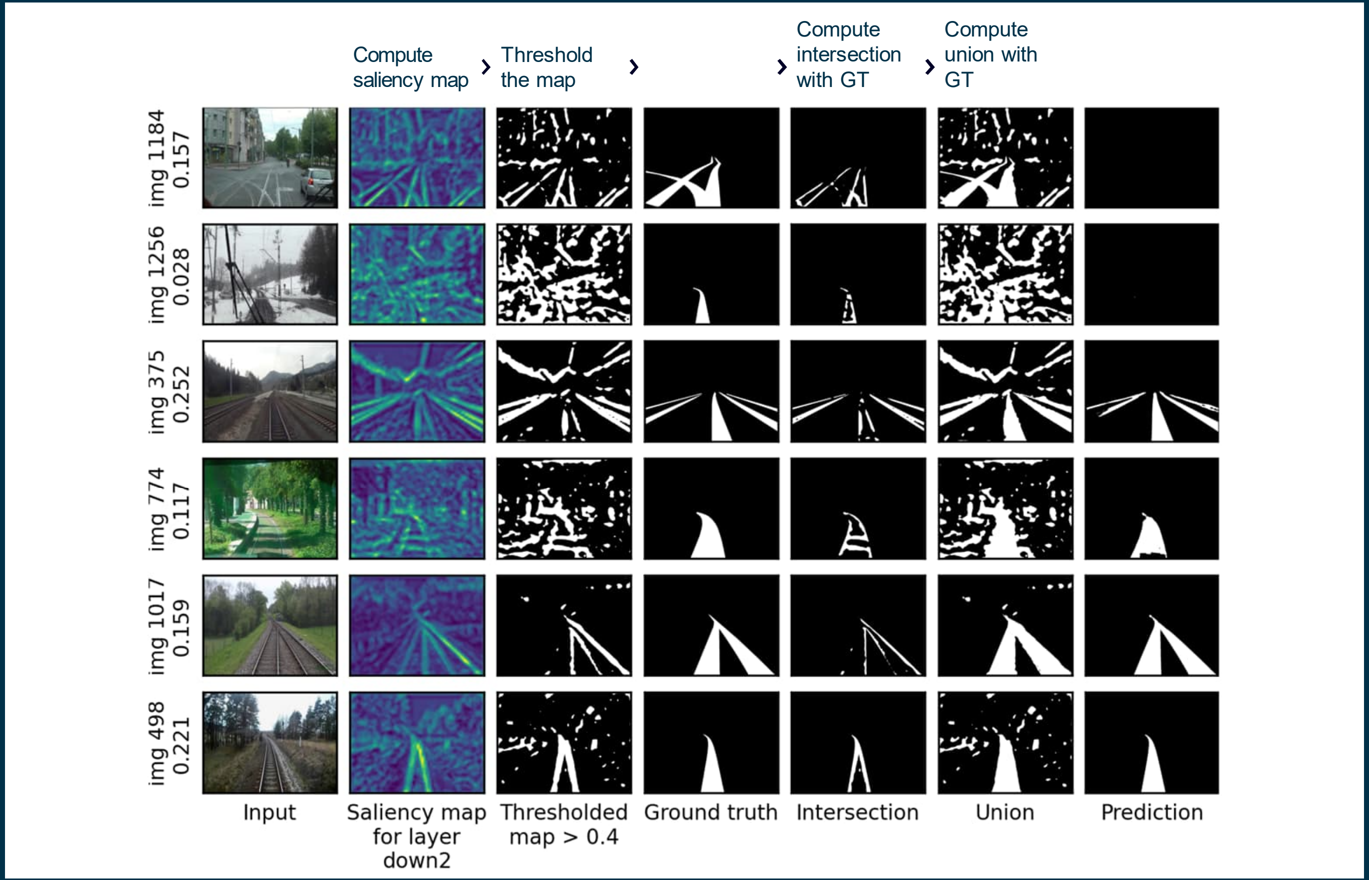
Identify “What portion of the network's ‘attention’ goes to the track when performing a track-segmentation task?”

Method

- Compute saliency map
- Threshold saliency map
- Compute intersection over union (IoU) between ground truth and saliency map

Purpose

- Compute visual similarity measure
- Provide a baseline for the future development
- Spark discussion
- Give tangible ground for exploring the applications of explainability methods for safety argumentation



Pillar 3: Concept-Based Explanations give insight into concept coverage and relevance, providing global explainability

Purpose: Even if the AI possess adequate performance, it must also be assessed that relevant concepts, e.g., of the ODD have been intrinsically encompassed by the system

Explain the model using high level human visual concepts (images). Concepts are both understandable and meaningful to humans

Globally explain the AI decision process with the underlying concepts, rather than the individual data points or parameters used in the model.

TCAV[1] score is calculated for each concept to know how relevant it is for the target class

Mitigation: Retrain the model with images containing the missing concept (tested)

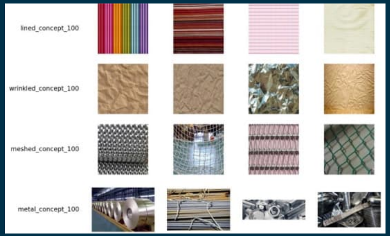
PROs

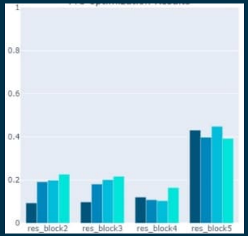
- Sole global method available
- increases transparency and trust into the model for a certifying party
- visual approach easy to comprehend by non data scientists

CONs

- Computationally very expensive
- Not all models support the necessary computations
- Missing clear guidelines for interpreting scores and setting reasonable thresholds

Basic concepts example: What concepts are relevant for track classification?

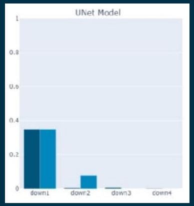


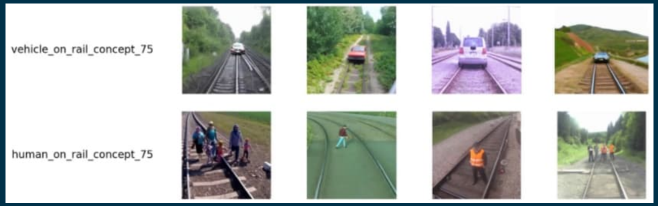


Result

- All concepts have been learned by the model

High-level concepts example: Have concepts vehicle/human on rail be learned?





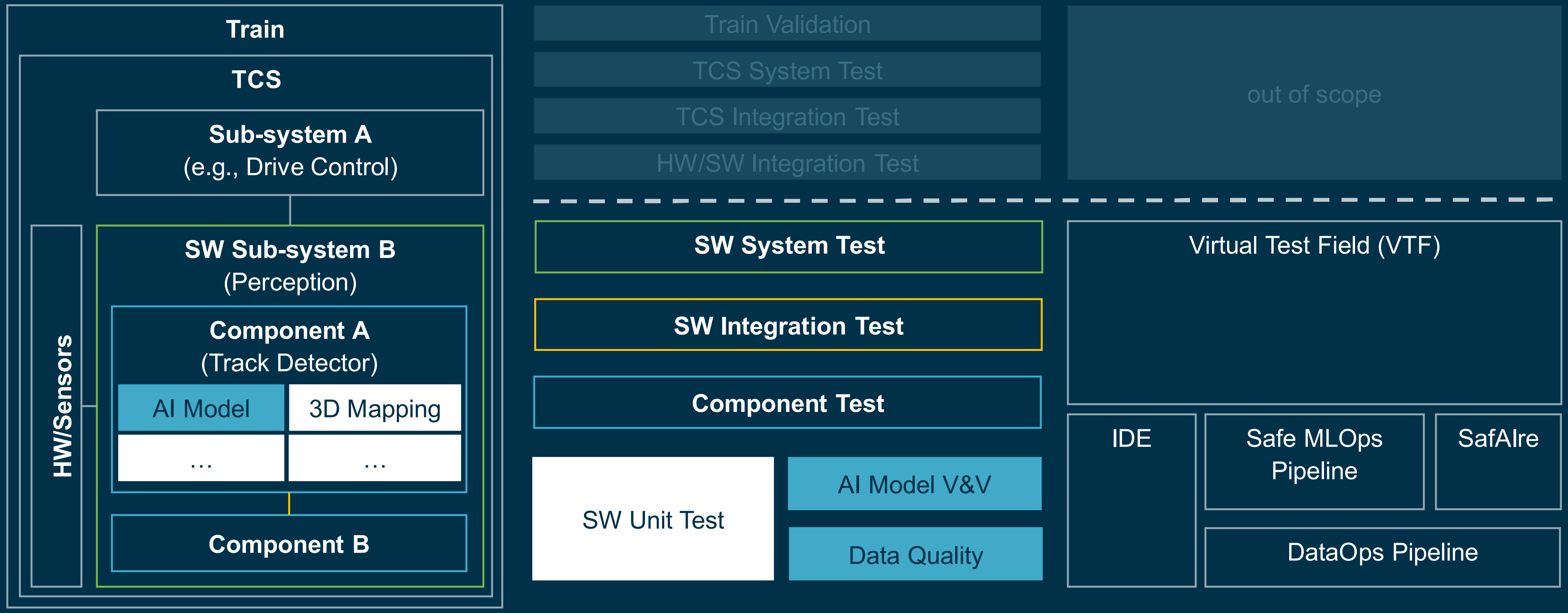
Result

- Both concepts for class “obstacle” in one of the layers
- Both have identical scores → “obstacle” class is paying attention to a common context

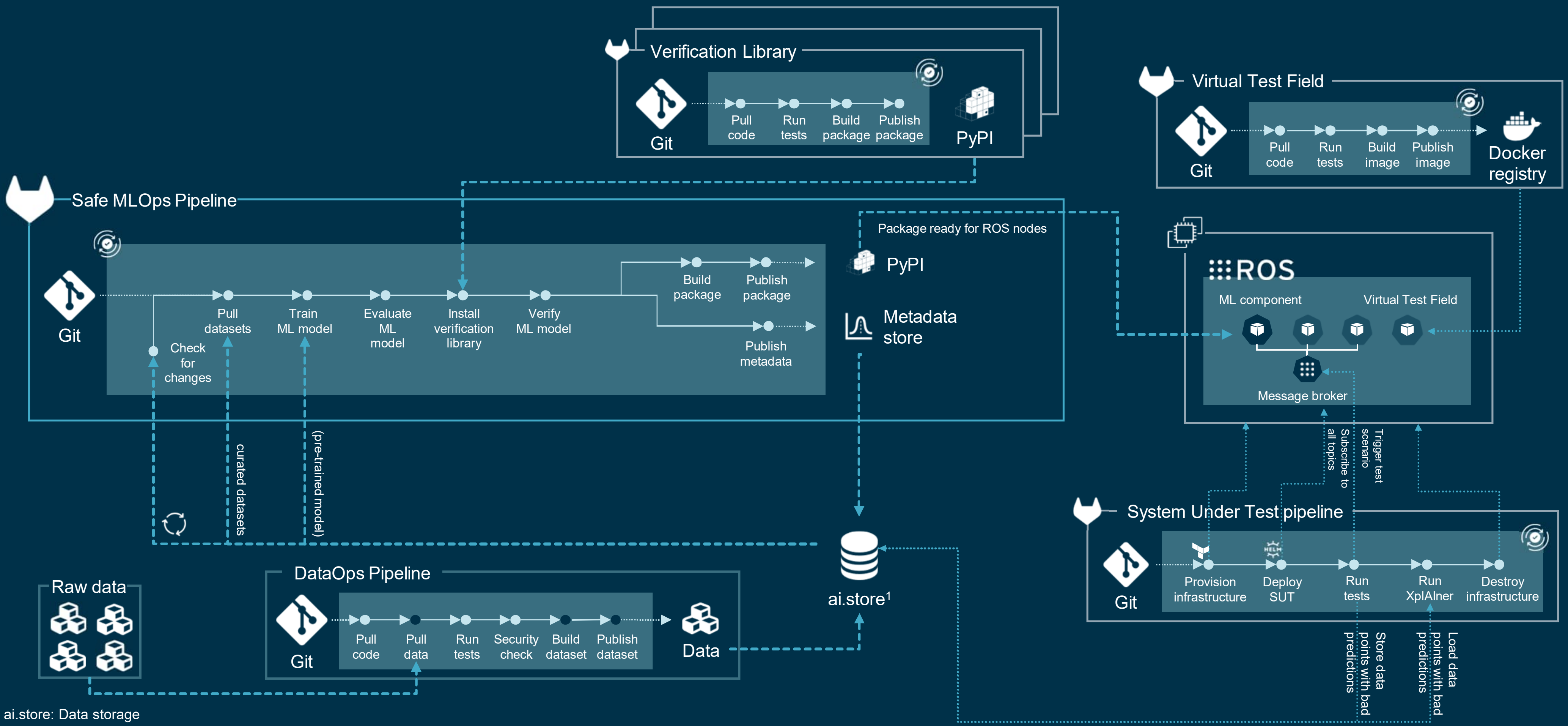
1 Kim et al 2022 “Quantitative Testing with Concept Activation Vectors (TCAV)”

Pillar 4: Each test level focuses on a specific test object and test goal and is supported by a corresponding test environment

Test Object (SUT) | Test Level | Test environment

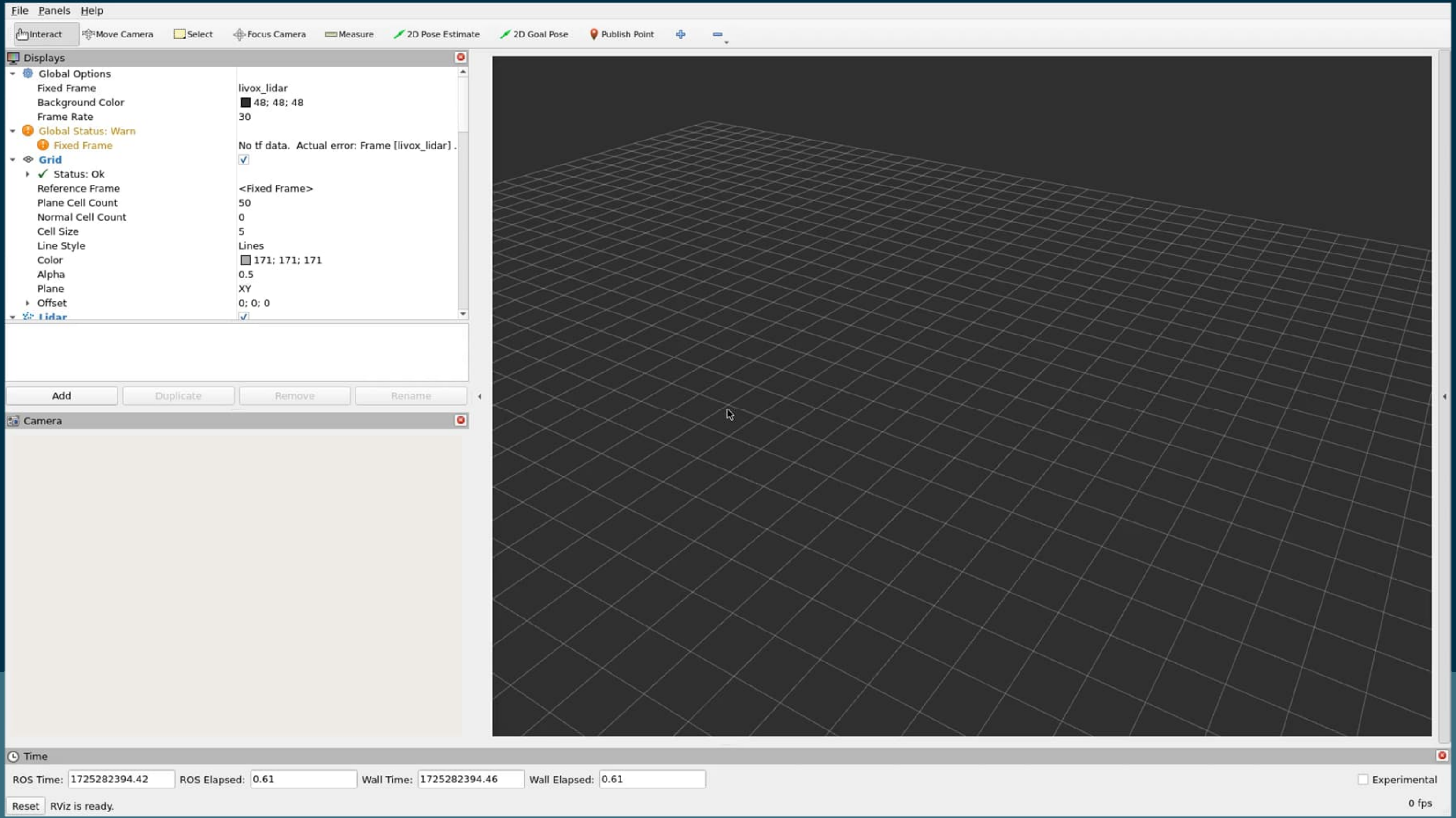


Pillar 4: Test environments in safe.trAIn



ai.store: Data storage

Pillar 4: For analysis of test results the VTF inputs and outputs are visualized



Pillar 5: Enhancing AI Safety through Runtime Monitoring of Out-of-Distribution Objects

Objectives

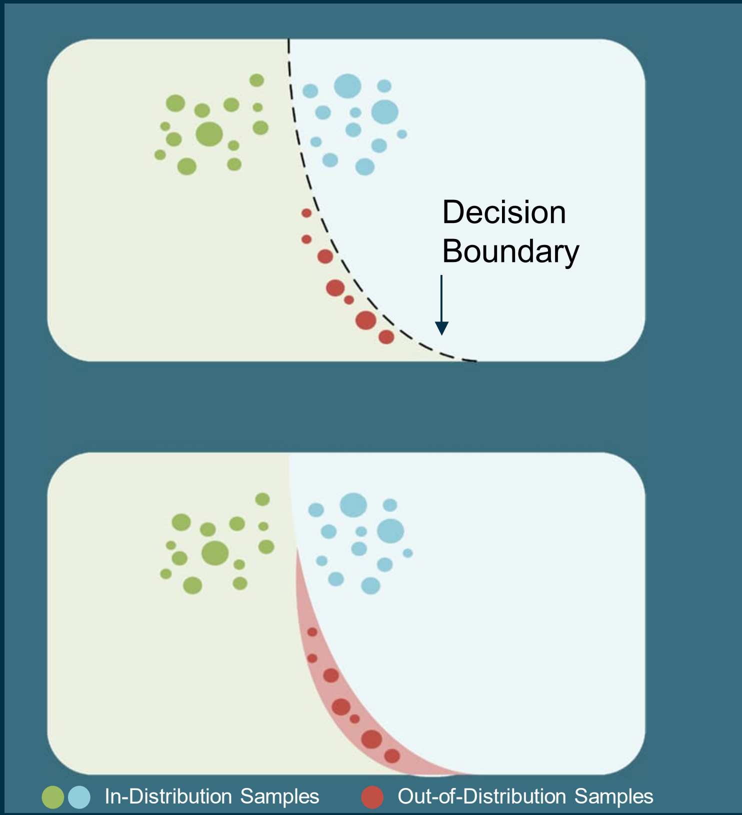
- Prevent unreliable AI model outputs when inputs deviate from the training distribution
- Ensure that the AI system adheres to specifications by monitoring its operation in real-time

Challenges

- Continuous monitoring introduces additional computational overhead, potentially impacting performance
- Distinction between valid OOD objects and background is challenging for widely varying sample distributions

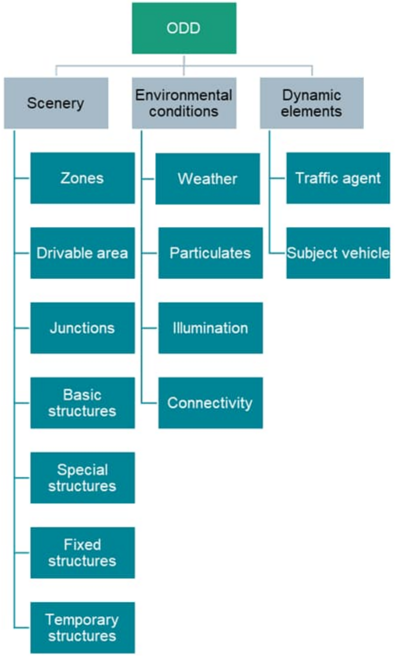
Approach

PROWL: A prototype-based zero-shot unsupervised OOD detection and segmentation framework



Pillar 5: How to Monitor Unknown Out-of-Distribution Elements

ODD



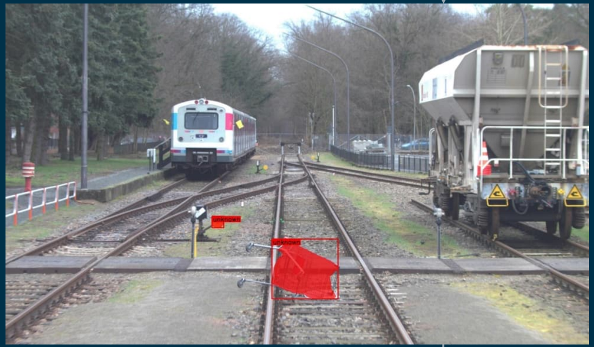
Out-of-Distribution

Elements that are **not** defined in the ODD are considered Out-of-Distribution (OOD).

PROWL | Prototype-based zero-shot unsupervised OOD detection and segmentation

- Relies on creating a prototype feature bank for each ODD object.
- Utilizes generalized robust features based on zero-shot inference with foundation model-based feature extractors

Example: Shopping Cart/Signal Box



PROWL correctly detects OOD objects like the shopping cart and the signal box which are not considered part of ODD in this setup.

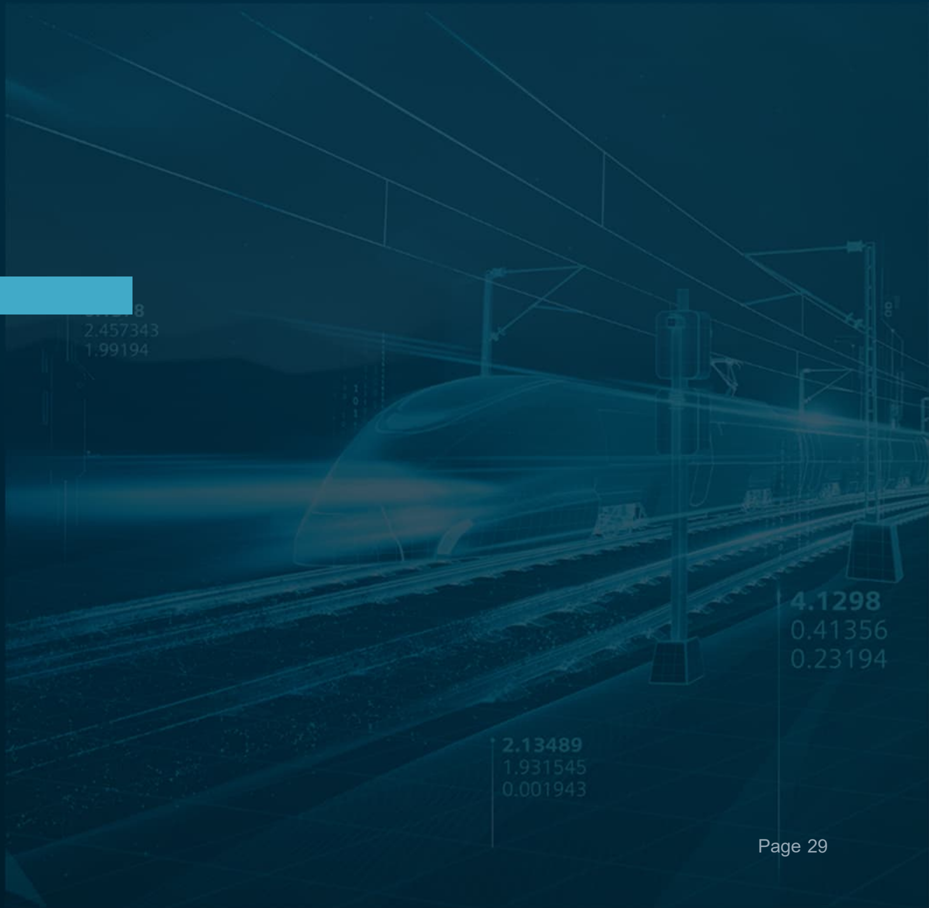
Example: Person Pose

Whenever significant features of ODD elements are not detected or visible, PROWL identifies them as (additional) OOD elements.

Sinhamahapatra, Poulami, et al. “Finding Dino: A plug-and-play framework for unsupervised detection of out-of-distribution objects using prototypes.” arXiv preprint https://arxiv.org/abs/2404.07664 (2024)

Summary & Outlook

15



Summary
safe.trAIn enables Safe Perception for Driverless Regional Trains

Challenges of AI in Railway

- No safety standard for AI-based perception in rail domain
- Unclear requirements for assessment of AI
- No established tools and processes

Project goals

Safe perception for automated trains

Safety-enabling architect-ture	Metrics/KPIs for (Self)-evaluation	Safety case and testing	Transfer to Standardi-zation

- Safety target approx. 1% Probability of Failure on Demand (PFD)
- 5 Pillars for safety assurance
  1. Processes
  2. Analysis of non-conventional redundancies
  3. Sufficient understanding of causalities
  4. Testing with real & simulated data
  5. Safety monitoring during operation
- Balance between the 5 pillars and how they can compensate for each other's weaknesses guides the safety validation
- “Landscape of AI safety concerns“ guides systematically the safety assurance
  - Analyzing ML-specific safety concerns
  - Find mitigating measures along the development life-cycle

Questions?

Dr. Marc Zeller

Siemens AG
Friedrich-Ludwig-Bauer-Str. 3
85748 Garching

marc.zeller@siemens.com

Funded by the European Union
NextGenerationEU

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

This research has received funding from the Federal Ministry for Economic Affairs and Climate Action (BMWK) under grant agreements 19I21039A.

safe·trAIn