# Safety Assurance of a Driverless Regional Train

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

Workshop “Critical Applications of AI in Industry, Healthcare and Other Sectors”

7th April 2025 | Marc Zeller, Siemens AG

# 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 (European AI ACT-high-risk application)
- 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

User

Technology Provider

Enabler

Assessor/ Standardization

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

## Safety objectives

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

## 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.

## 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.”

# The overall safety target relates to the concept of Recall

According to CSM RA “comparison with reference system”

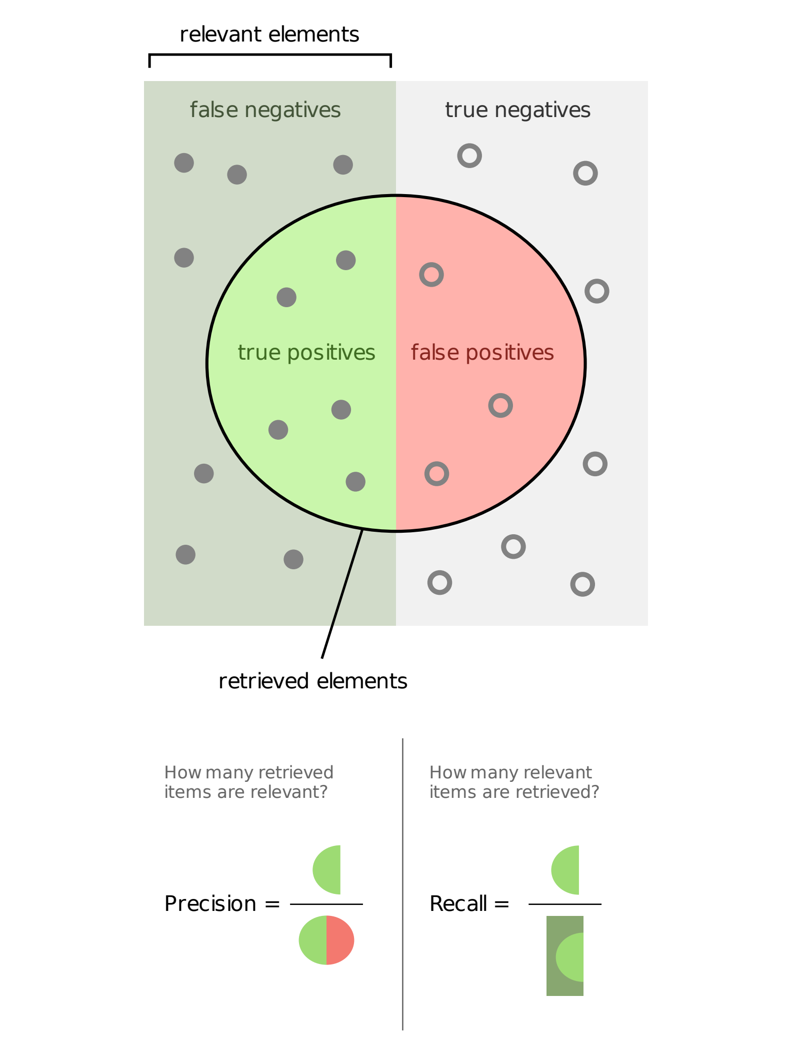
## Safety target: “overall as good as driver”

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

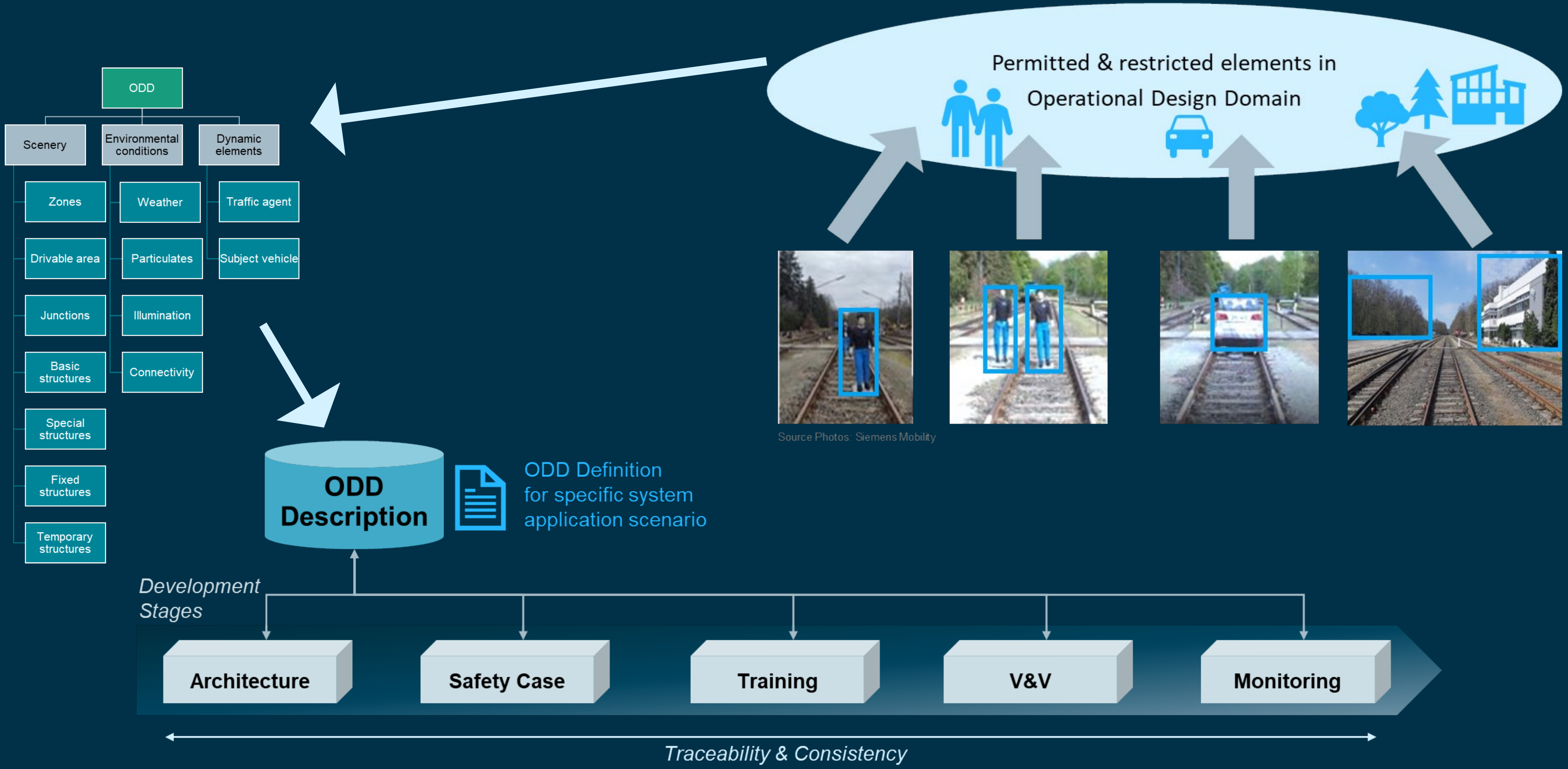
safe·trAIn

**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]:

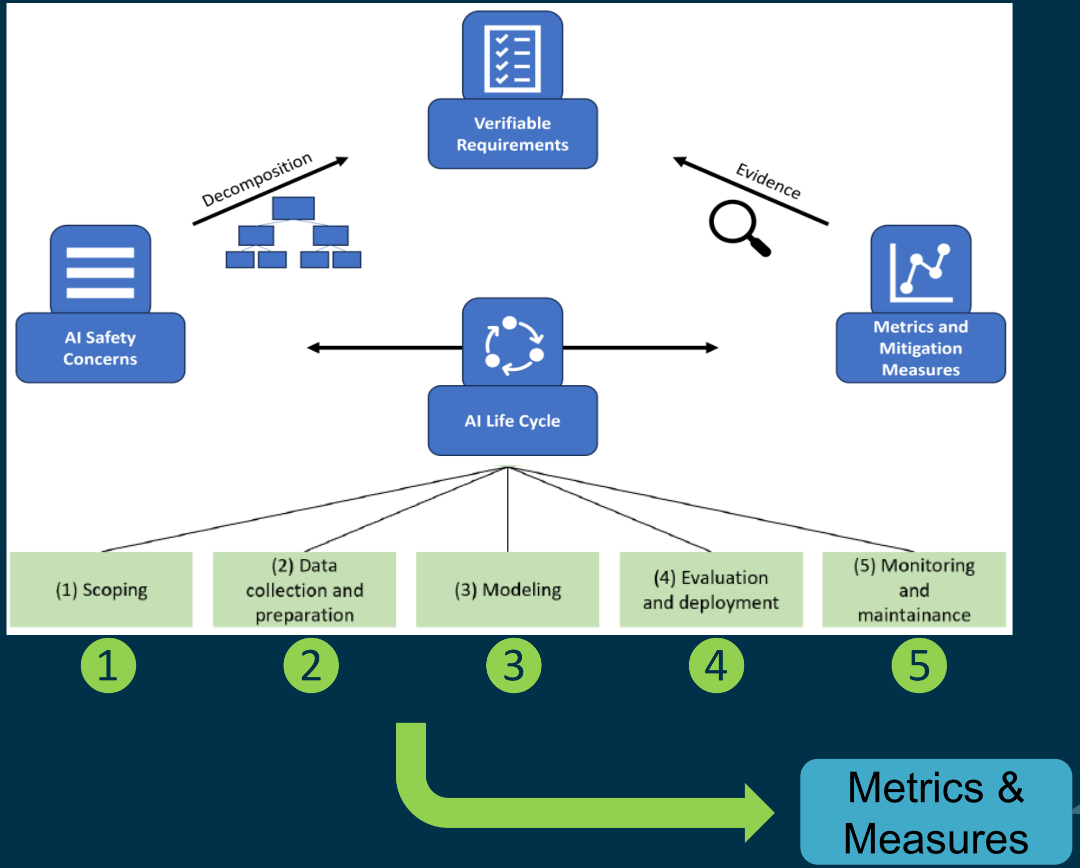
## AI Safety Concerns[1]

| | | | | | | |
|---|---|---|---|---|---|---|
| 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 |

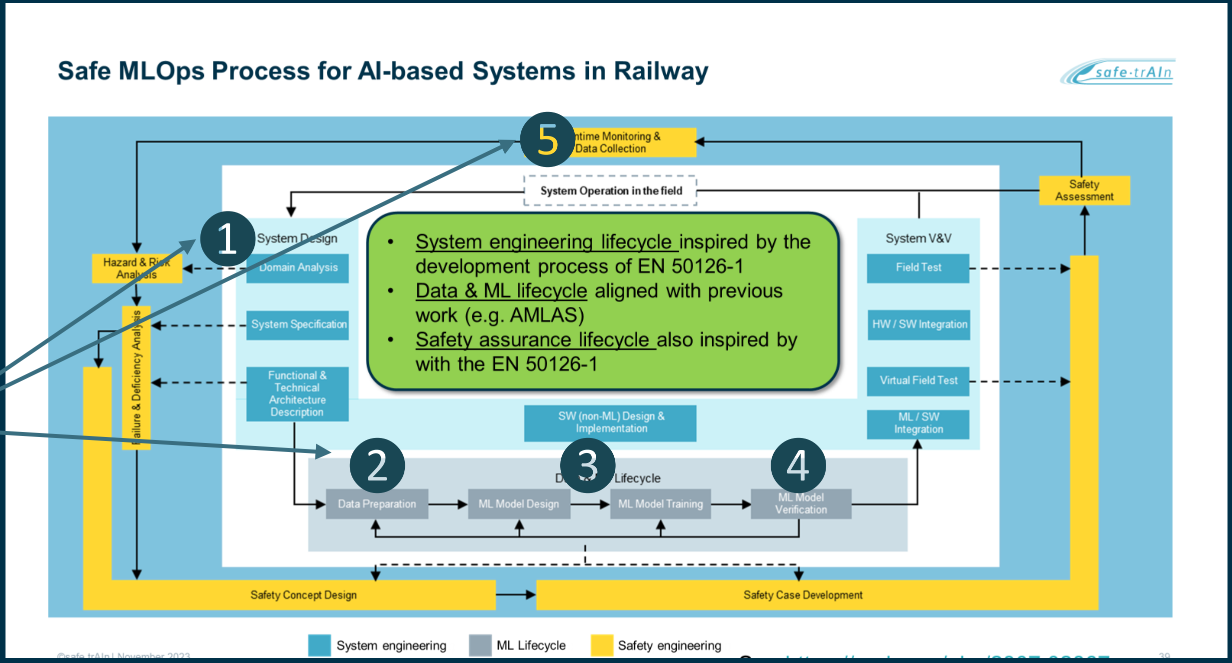
**1** Schnitzer, R., Hapfelmeier, A., Gaube, S., Zillner, S.: AI Hazard Management: A framework for the systematic management of root causes for AI risks. | **2** Houben, S., Abrecht, S., Akila, M., Bär, A., Brockherde, F., Feifel, P., et al.: Inspect, Understand, Overcome: A Survey of Practical Methods 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: Landscape of AI Safety Concerns and safe MLOps Process




In order to assure AI-based autonomous systems:

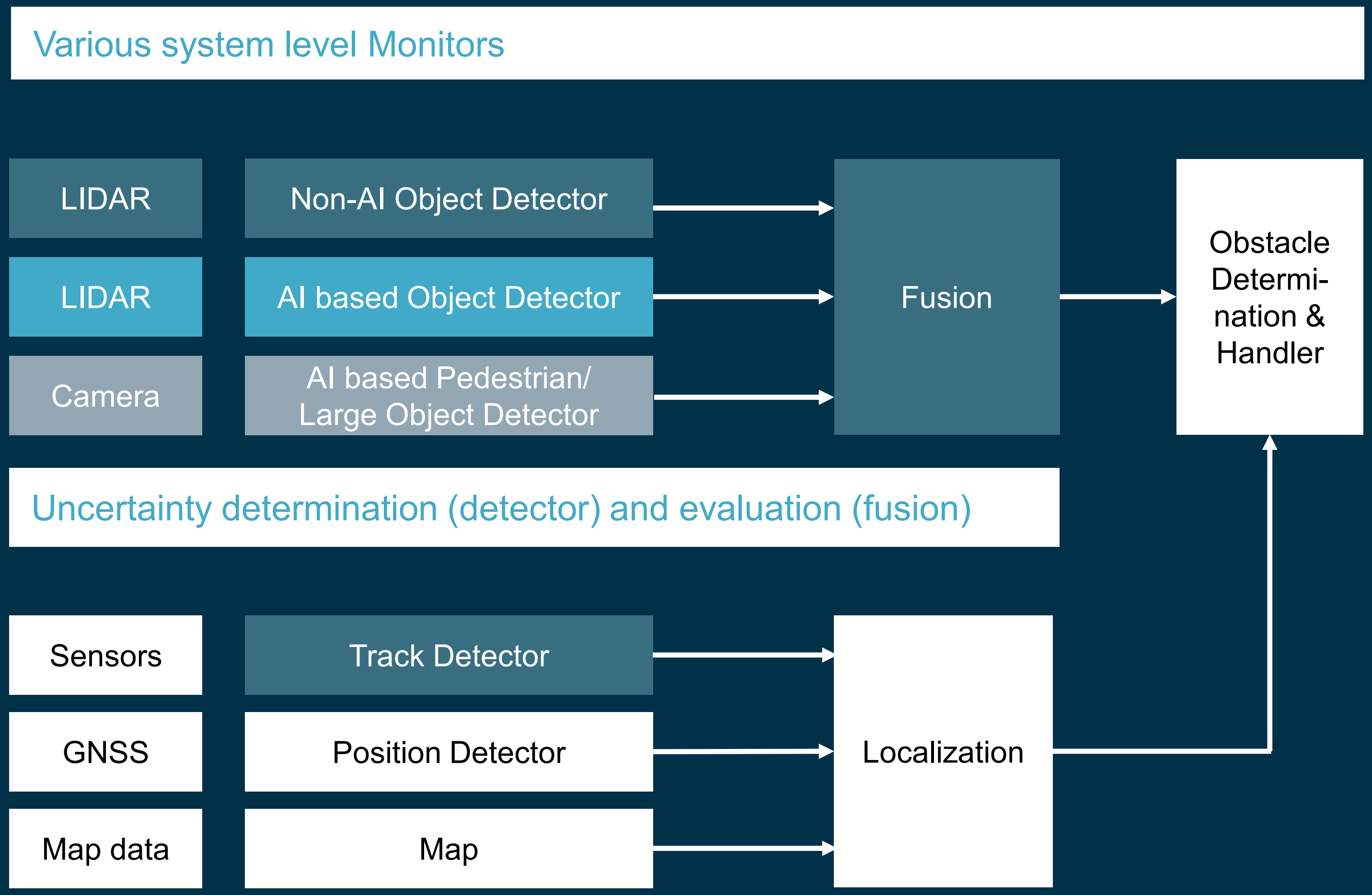
For each **AI Safety Concern, evidence** needs to be derived along **the whole AI life cycle** that **convincingly demonstrates** the sufficient mitigation of the respective AI Safety Concern.




**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**

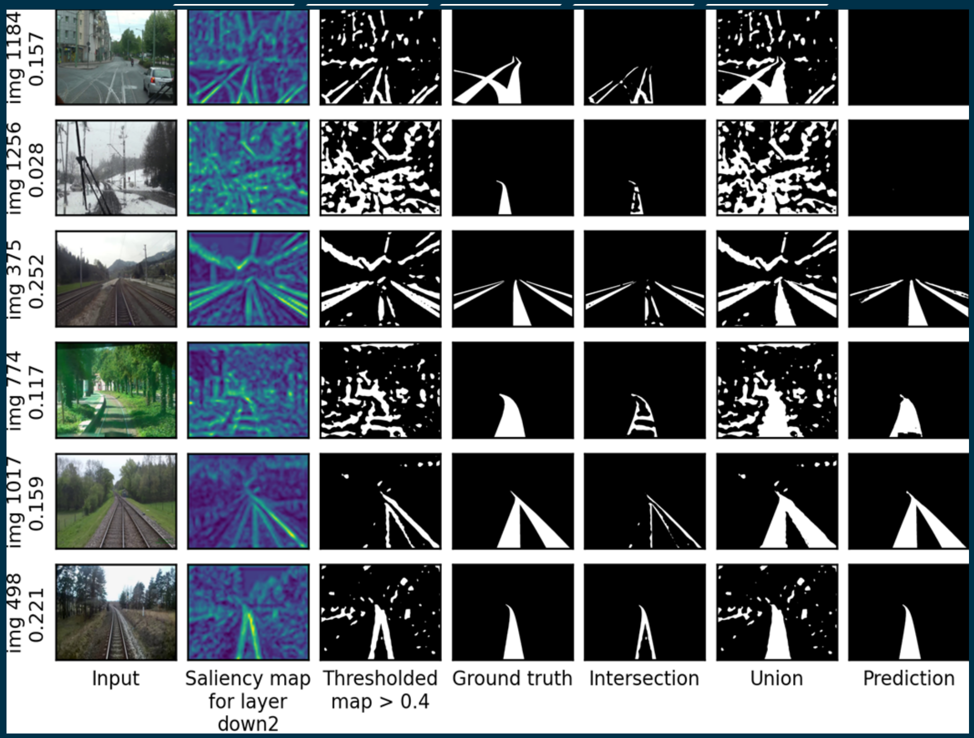
---

**Monitoring of system and components at runtime**

- Safety measures realized in monitors and components

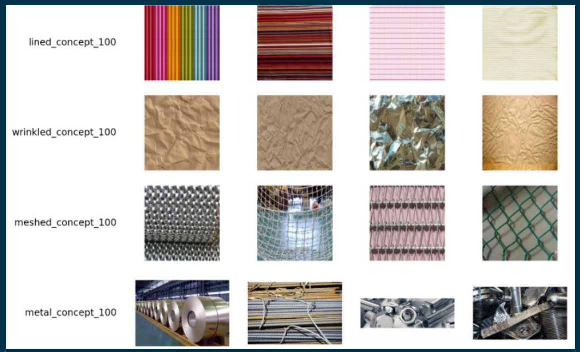
# Pillar 3: Sufficient Understanding of Causalities using eXplainable AI

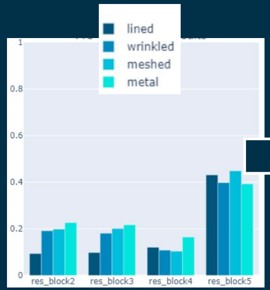
- Saliency map is a 2D image which shows the most important regions on the input image
- Possible meaning of the metric: "What portion of the network's "attention" goes to?




- Explain the model using high level human (visual) concepts
- Globally explain the AI decision process with the underlying concepts using TCAV approach

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




**Result**

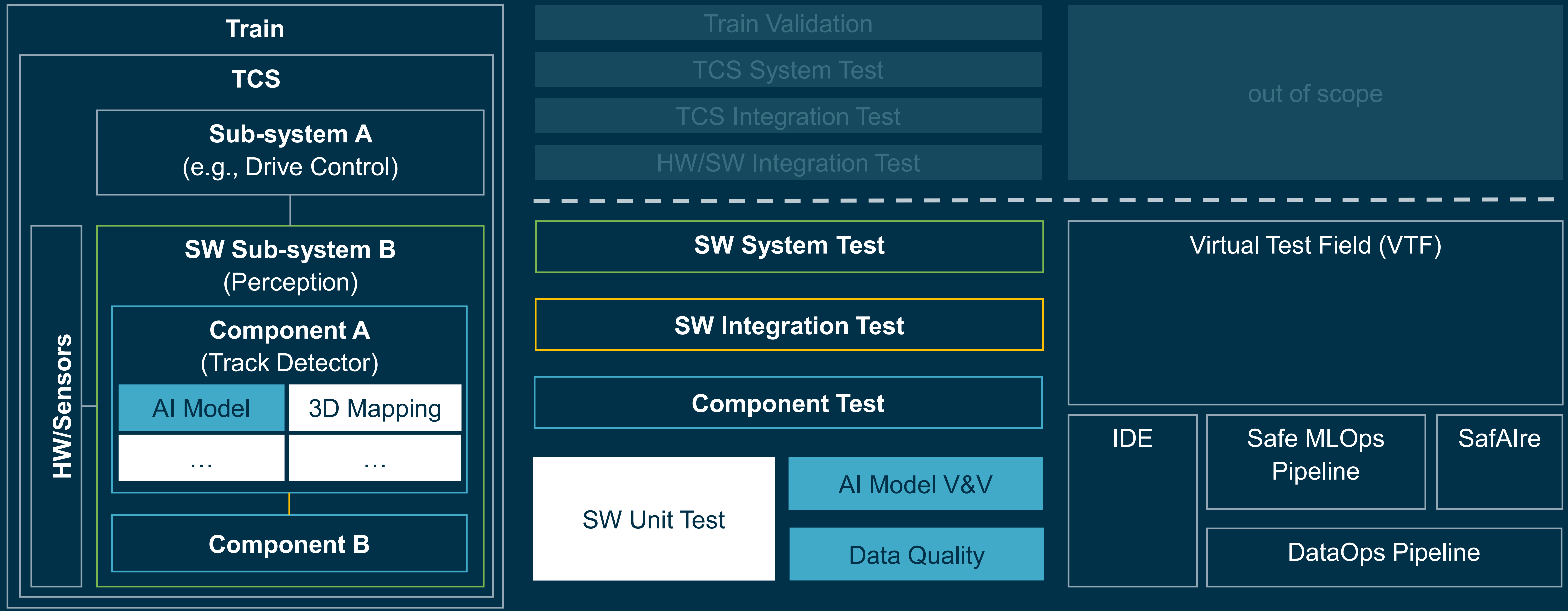
- All concepts have been learned by the model

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). InInternational conference on machine learning 2018 Jul 3 (pp. 2668-2677). PMLR.

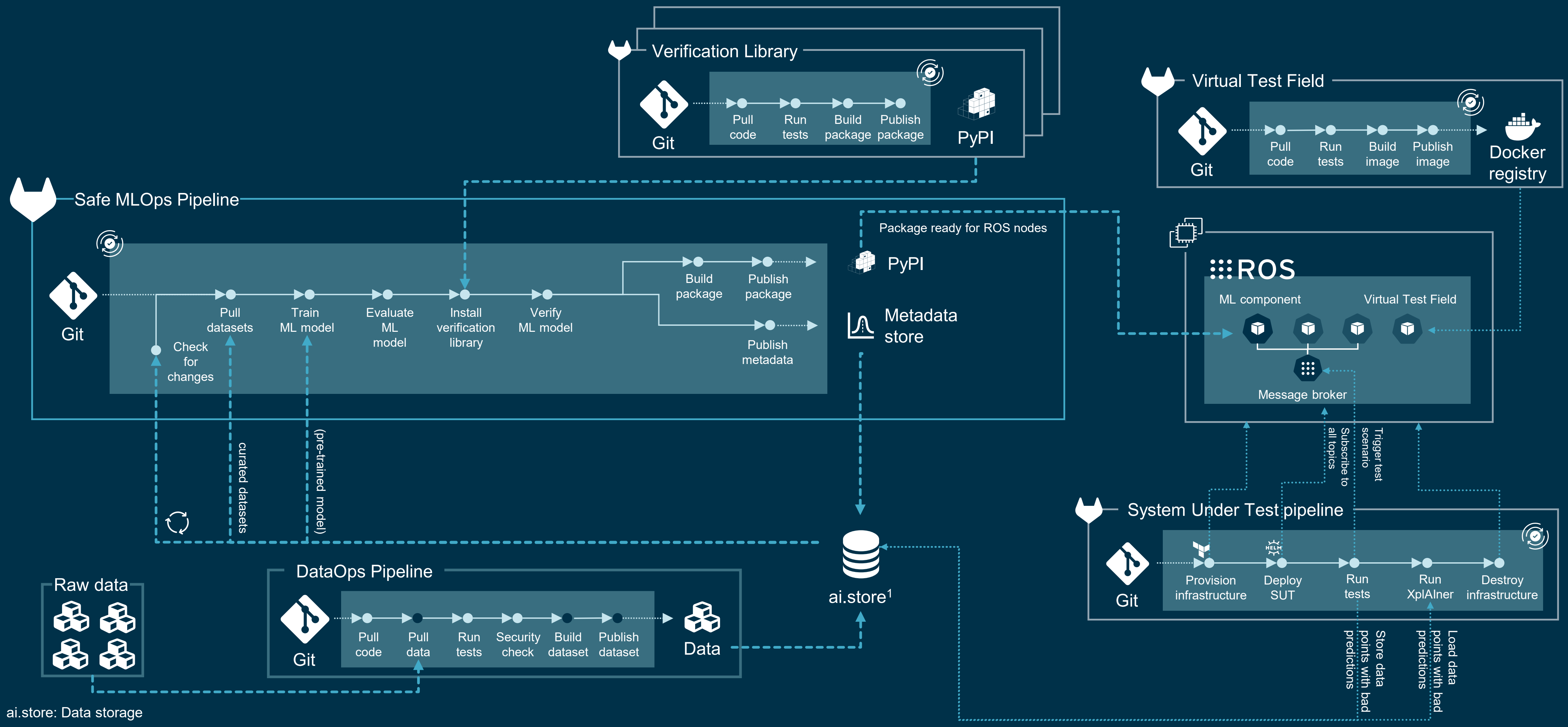
# Pillar 4: Each test level focuses on a specific test object and test goal and is supported by a corresponding test environment

safe·trAIn

**Test Object (SUT)** | **Test Level** | **Test environment**

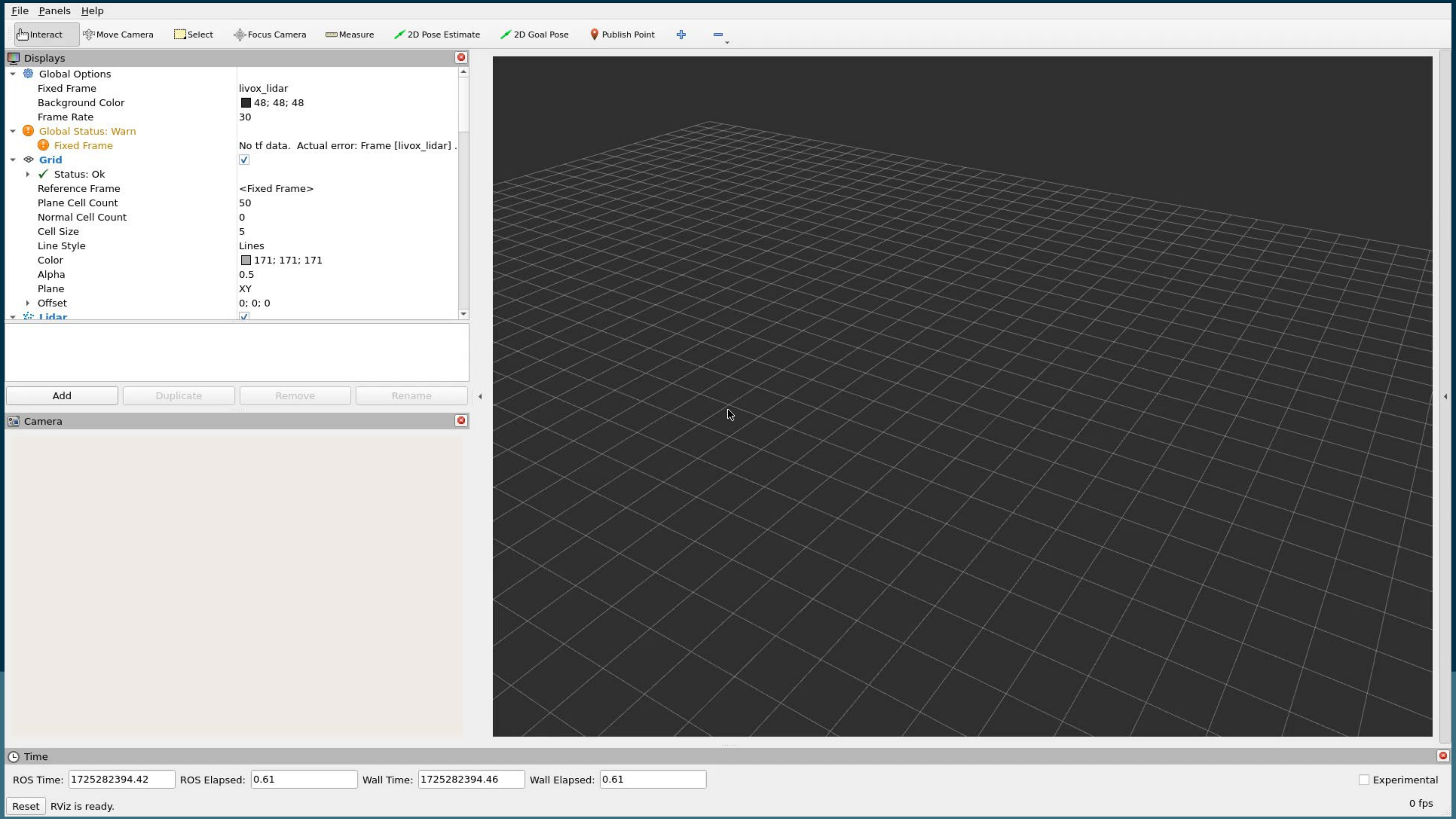


# Pillar 4: Test environments in safe.trAIn




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

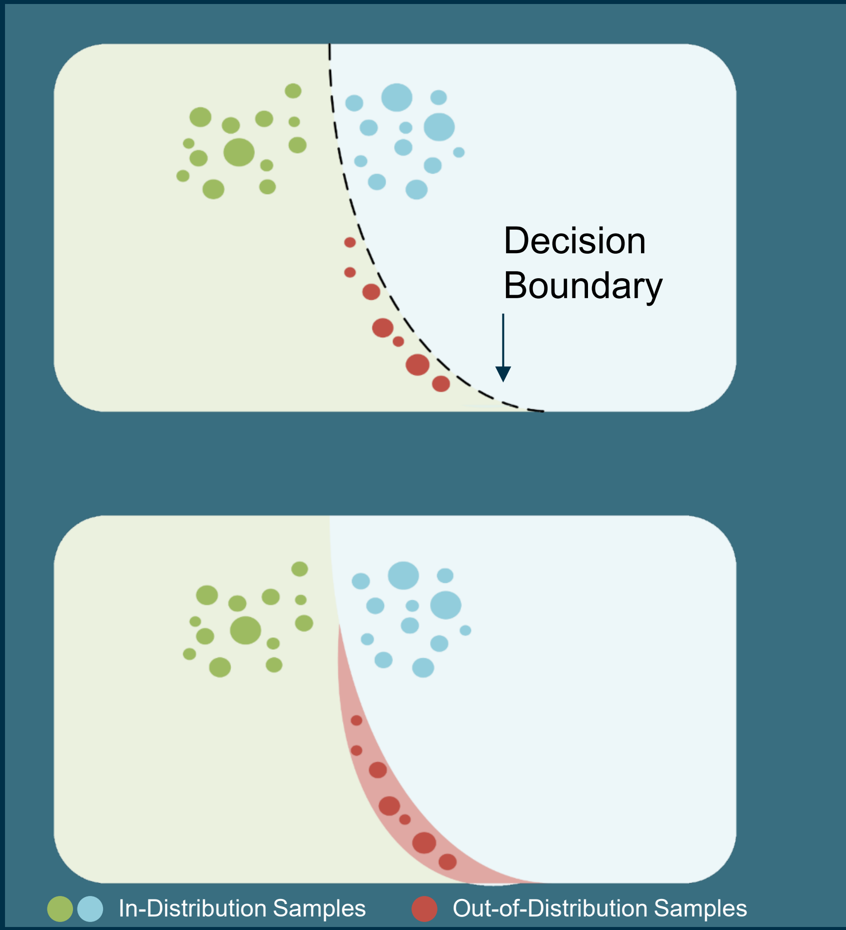
- Prevent unreliable AI model outputs when inputs deviate from the training distribution
- Ensure that the AI system adheres to specifications by monitoring its operation in real-time

## Challenges

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

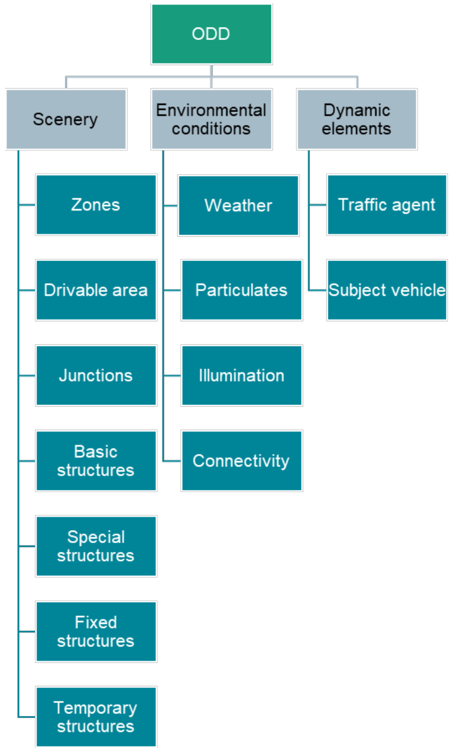
## Approach

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



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

## ODD



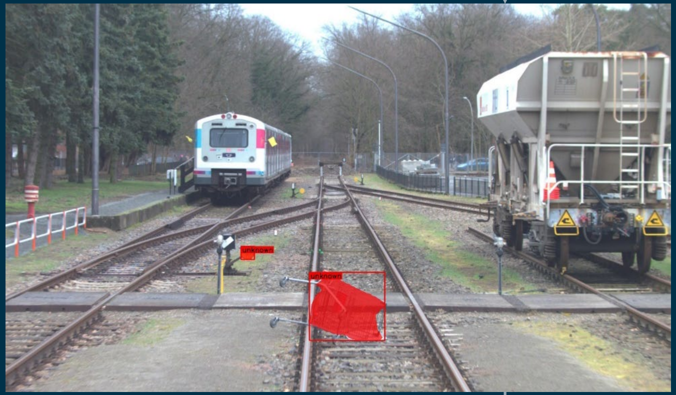

### Out-of-Distribution

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

safe·trAIn

### Challenges of AI in Railway

- No safety standard for AI-based perception in rail domain
- Unclear requirements for assessment of AI (European AI ACT- high-risk application)
- 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 |
|---|---|---|---|

**Next Steps:**
- **Transfer to international standardization**
- **Follow up projects towards products**

- 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

# Outlook

## Transfer of safe.trAIn results to other domains

### 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

- AI Safety Concerns are domain and use case independent
- Tailoring to specific use cases is required
- Application to robotic use cases currently done in the RoX project

safe·trAIn

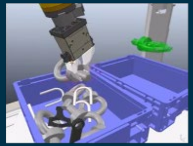
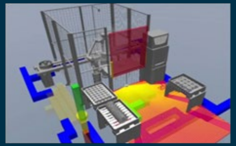
Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
Finanziert von der Europäischen Union
NextGenerationEU
aufgrund eines Beschlusses des Deutschen Bundestages

ROX
Enabling AI Robotics







# 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