

Rail to Digital automated up to autonomous train operation

D6.4 – Proposal for the Certification Process of Perception Systems

Due date of deliverable: 30/11/2025

Actual submission date: 30/11/2025

Leader/Responsible of this Deliverable: Dr. Thomas Waschulzik

Reviewed: Y

Document status		
Revision	Date	Description
01	15/08/2025	First major version
02	02/10/2025	Initial version of all major chapters integrated
03	06/10/2025	First review comments resolved
04	17/10/2025	Final version for TMT Review
05	7/11/2025	Version after the integration of TMT feedback
06	17/12/2025	Version after the integration of SteCo feedback
07	29/09/2026	Version after the integration of MCP 10 feedback

Project funded from the European Union's Horizon Europe research and innovation programme		
Dissemination Level		
PU	Public	X
SEN	Sensitiv – limited under the conditions of the Grant Agreement	

Start date: 01/11/2023

Duration: 24 months

ACKNOWLEDGEMENTS



This project has received funding from the Europe's Rail Joint Undertaking (ERJU) under the Grant Agreement no. 101102001. The JU receives support from the European Union's Horizon Europe research and innovation programme and the Europe's Rail JU members other than the Union.

REPORT CONTRIBUTORS

Name	Company	Details of Contribution
Thomas Waschulzik	SMO	First version / General structure Executive Summary, First version of Chapter 1, 2, 3.3, 3.4, 3.6, 3.7, 3.8, 5, 6, 8
Hélène ARFAOUI KAYNAK	SNCF	Contributions to state of art overview on DEEL, the regulatory framework, and results of the PER safety analysis, improvements and limitations and deliverable review
Quentin BLERIOT	ALSTOM	Contributions to summary of state of art (IEC 61508/TR 5469)
Andrea BELLITTO	WABTEC	Contribution to 4.2 Requirements for the perception system, 4.3 Architecture of the perception system, and deliverable review
Dominik KEVICKÝ	AŽD	Reviewer

Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the author’s view – the Joint Undertaking is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability.

EXECUTIVE SUMMARY

General

Deliverable D6.4 specifies a certification process for AI-enabled Perception systems (PER) on rolling stock within Flagship Project 2 (R2DATO), aligning safety, interoperability and regulatory obligations for deployment in European rail operations. It consolidates foundations from safety and AI standards, maps them to railway regulations and proposes a general and extended workflow (Stages 0–9) with clearly delineated roles for the European Union Agency for Railways (ERA), National Safety Authorities (NSAs), Notified Bodies (NoBos), Designated Bodies (DeBos), independent Assessment Bodies (AsBos) and stakeholders (IMs, RU, Suppliers). The approach interfaces with Technical Specifications for Interoperability (TSIs)—notably LOC&PAS, CCS and OPE—and the Common Safety Method for Risk Evaluation and Assessment (CSMRA), the recommendations from EN 50716 regarding AI, and considers the EU Artificial Intelligence Act obligations for high-risk AI systems. Dependencies to WP6.3 (architecture and system specification) and WP8 (safety analysis) are made explicit. The document defines responsibilities for AI-related artefacts that are required to bring a PER system into operation including Operational Design Domain (ODD), datasets, HD maps, simulation fidelity, uncertainty reporting.

Scientific/Technical Approach

The D6.4 considers as normative state-of-the-art for voluntary application IEC 61508 (demand modes, SIL allocation), ISO/IEC TR 5469:2024 (AI safety guidance across data, learning and inference), DIN DKE SPEC 99002 (terminology) and DIN DKE SPEC 99004 (rail ODD taxonomy). State-of-the-art inputs from projects TAURO D1.2, DEEL, Confiance.ai (Body of Knowledge) and safe.trAIIn are synthesised to adapt system-engineering “V” lifecycles with ML-specific “W” learning assurance, scenario-based verification and layered safety cases. The workflow spans ODD definition, HD-map preparation, simulation, test with hardware and software in the loop, risk analysis under CSMRA, verification & validation (lab, simulation and field), conformity assessment (TSIs and national rules), independent safety assessment and authorisation. ERA/NSA roles and NoBo/DeBo/AsBo interfaces are embedded throughout. The document records differing positions on PER demand mode (continuous vs. Low-demand) and links Probability of Failure on Demand (PFD) to ML recall for acceptance, while noting EN 50716’s caution that AI/ML remains “not recommended” for SIL1–SIL4 (basic integrity only), requiring compensating architectural and process measures.

Key Findings

- For the analysed use cases under European Train Control System (ETCS) Full Supervision, WP8 indicates Safety Integrity Level (SIL) needs up to SIL2 for certain PER safety-relevant functions; applicability depends on assumptions, external barriers and project context.
- Implementing a perception system for driverless trains is complex. New artefact responsibilities, stakeholder skills, and standards must be defined before it is possible to introduce GoA3/4 systems efficiently to the market.
- A modular ODD, with machine-readable descriptors, improves traceability and reuse across projects; acceptance metrics must cover detection and false-alarm characteristics, latency, availability and uncertainty, with mapping to safety goals (e.g., recall→PFD).
- ODD definition, dataset and labelling governance (provenance, representativeness, version control, GDPR compliance) and HD-map quality/change control are essential parts of the assurance process.
- Simulation fidelity (infrastructure, environment, sensors), correlation with field data and structured scenario libraries are necessary to generate scalable evidence.

- Standardisation is evolving (CEN/CENELEC JTC 21 programme, ISO/IEC workstreams); alignment activities are required to resolve current gaps between AI guidance and rail-specific safety norms. The standardization activities supporting the EU AI ACT are late, resulting in uncertainties on how to assure compliance with the EU AI ACT for high-risk applications.

Recommendations and Future Steps

Prioritise on further clarifications of the stakeholder responsibilities based on the proposed starting point in this document will be important to enable successful execution projects implementing GoA3/4 systems in open environment. The responsibilities in D6.4 Process export constraints and contract-based cooperations which have been considered too stringent by some stakeholders. This document can be used as a starting point for the future work.

Further continued standardization activities will support the content and quality of new required artefacts like the ODD definition, HD-maps, simulations supporting the training, test and validation of perception systems, as well as metrics measuring the important aspects of a PER system.

ABBREVIATIONS AND ACRONYMS

ADM	Automatic Driving Module
AI	Artificial Intelligence
APM	Automatic Processing Module
ATO	Automatic Train Operation
ATP	Automatic Train Protection
BAMS	Brake-Adhesion Management System
BG	Balise Group
BoK	Body of Knowledge (from Confiance.AI)
CCS	Control, Command and Signalling
CCTV	Closed circuit television
CLC	European Committee for Electrotechnical Standardization
CNNS	Convolutional Neural Networks
DEEL	Dependable, Explainable & Embeddable Learning (Name of a french R&D Project)
DMU	Diesel Multiple Unit
EB	Emergency Brake
ECU	Electronic Control Unit
EMC	Electromagnetic Compatibility
EMU	Electric Multiple Unit
EN	European Standard
EoA	End of Authority
ERA	European Union Agency for Railways
ERJU	Europe's Rail Joint Undertaking

ERTMS	European Rail Traffic Management System
ETCS L1 LS	European Train Control System - Level 1 – Limited Supervision
ETCS	European Train Control System
ETCS-OB	European Train Control System – Onboard
EU	European Union
EVC	European Vital Computer
FMEA	Failure Modes and Effects Analysis
FMS	Fleet Management System
FP	Flagship Project
GoA	Grade of Automation
hEN	Harmonized European standards
HiL	Hardware in the Loop
HRI	High-Risk AI
IEM	Emergency Manager
IIM	Incident Manager
IM	Infrastructure Manager
IMS	Incident Management System
IOM	IM Operations Manager
IPM	Incident and Prevention Manager
IPM-ISM	Incident and Prevention Manager – Incident Solving Manager
IPM-OB	Incident and Prevention Manager – Onboard
IPX	Innovation Programme X (Shift2Rail)
IT	Information Technology
JP	Journey Profile
LAISC	Landscape of AI Safety Concerns (LAISC)

LC	Level Crossing
LOC	Localization
LOC & PAS TSI	Rolling Stock – Locomotives and Passengers Technical Specification for Interoperability
MA	Movement Authority
ML	Machine Learning
NLF	New Legislative Framework
OAS	Onboard Automation System
OCC	Operations Control Center
ODD	Operational Design Domain
OJ	Official Journal
OM	Operations Manager
OMTS	Onboard Multimedia and Telematic Subsystem
OPE / TSI OPE	Operation and Traffic Management Technical Specification for Interoperability
OS mode	On sight mode
OSS	ERA operates the OneStopShop (OSS) platform
PER	Perception
PIS	Passenger Information System
PTU	Physical Train Unit
R2DATO	Rail to Digital automated up to Autonomous Train Operation
RAMS	Reliability, Availability, Maintainability, and Serviceability
REP	Repository
RO	Railway Operator
RU	Railway Undertaking
RUS	Railway Undertaking Supervisor

SCV	Signal Converter
SiL	Software in the Loop
S2R	Shift2Rail Joint Undertaking
SIL	Safety Integrity Level
SRS	System Requirements Specification
SL	Line side signalling
TAS	Trackside Automation System
TAURO	Technologies for Autonomous Rail Operation
TCMS	Train Control Management System
TE	Technical Enabler
TM	Traffic management
TMS	Traffic Management System
TR	Technical Report
TRL	Technology Readiness Level
TS	Train system
TSI CCS	Control, Command and Signalling Technical Specification for Interoperability
TSR	Temporary Speed Restriction
UC	Use Case
UIC	Union Internationale des Chemins de fer (International Union of Railway)
WP	Work Package
XiL	X=H, S Hardware or Software in the Loop
X2Rail-4	Advanced signalling and automation system – Completion of activities for enhanced automation systems, train integrity, traffic management evolution and smart object controllers

Acknowledgements.....	2
Report Contributors.....	2
Executive Summary	4
Abbreviations and Acronyms.....	6
1 Introduction	13
2 Goals of WP 6.4	15
3 Summary of the state of art	16
3.1 IEC 61508-1:2010 Functional safety	16
3.2 ISO TR 5469:2024	18
3.3 DIN DKE SPEC 99002.....	21
3.4 DIN DKE SPEC 99004.....	22
3.5 DEEL project.....	25
3.5.1 Challenges of ML in Certified Systems	25
3.5.2 Certification practices in the railway industry.....	26
3.5.3 Towards new certification practices in the railway sector	27
3.5.4 Towards certification of ML	27
3.5.5 Opportunities for ML in industry	28
3.5.6 DEEL way-forward.....	28
3.6 TAURO D1.2 – Proposal for certification methodology of safe functions based on AI	29
3.6.1 Introduction and Context.....	29
3.6.2 Normative and Regulatory Framework.....	29
3.6.3 Trustworthy AI: Principles and Dimensions	30
3.6.4 Development Methodology: V-Model, Scenario Abstraction and Safety	31
3.6.5 PEGASUS Safety Argumentation and Layered Structure	31
3.6.6 Data Management and Traceability	32
3.6.7 Test Environments and Validation Methods	33
3.6.8 Evidence and Acceptance Criteria	34
3.6.9 Test Data Management and Harmonisation	36
3.6.10 Recommendations and Future Steps.....	37
3.6.11 Conclusion.....	37
3.7 The community to accelerate the deployment of responsible AI in industrial systems	38
3.7.1 Introduction and overview of Confiance.ai.....	38
3.7.2 The Body of Knowledge (BoK).....	39
3.7.3 Activities for engineering a critical ML-based system	41
3.7.4 The Confiance.ai knowledge graph.....	42

3.7.5	Implications for WP6.4 and recommendations	43
3.7.6	Conclusion	44
3.8	Safe AI using a driverless regional train as an example	44
3.8.1	Purpose of the safe.trAI demonstration	44
3.8.2	Acceptance criteria for the safety function	45
3.8.3	Landscape of AI Safety Concerns (LAISC)	46
3.8.4	Safe perception system architecture	49
3.8.5	Safety-case strategy for the GoA3/4 architecture	51
3.8.6	Lessons for R2DATO WP 6.4	54
4	Input in the PProcess Description	56
4.1	Regulatory and normative Framework	56
4.1.1	Technical Specification for Interoperability	56
4.1.2	Common safety Methods Risk assessment	57
4.1.3	EU AI ACT	57
4.1.4	EN 50716	62
4.2	Requirements for the Perception system	63
4.2.1	Functional Requirements	63
4.2.2	Non-Functional requirements – Real-Time Safety	64
4.3	Architecture of the perception system	64
4.3.1	General approach to the architecture	64
4.3.2	High level description of the architecture	64
4.4	Safety requirements	65
5	Certification Process Description	69
5.1	Actors in Force	69
5.1.1	European and national authorities	69
5.1.2	Industrial stakeholders	70
5.1.3	Interaction overview	71
5.2	General Certification Workflow Overview	72
5.2.1	Concept and scoping	72
5.2.2	Requirements and architecture alignment	72
5.2.3	Data foundations	73
5.2.4	Safety and risk management	73
5.2.5	Verification and validation strategy	73
5.2.6	Conformity assessment	74
5.2.7	Safety assessment	74

5.2.8	Authorization and deployment.....	74
5.3	Special Aspects and Requirements for Perception Systems for the Certification	74
5.3.1	Modular operational design domain (ODD)	75
5.3.2	Performance and acceptance metrics	75
5.3.3	Dataset and labelling governance	76
5.3.4	HD map quality and change management	77
5.3.5	Simulation fidelity	77
5.3.6	Interfaces and interoperability	78
5.3.7	Safety case evidence map	78
5.4	Extended Workflow Overview	78
5.4.1	Stage 0 – Initiation and governance	79
5.4.2	Stage 1 – ODD baseline definition	79
5.4.3	Stage 2 – HD map definition	80
5.4.4	Stage 3 – Simulation environment preparation.....	81
5.4.5	Stage 4 – Safety, hazard analysis and architecture	81
5.4.6	Stage 5 – Verification (lab and simulation tests).....	82
5.4.7	Stage 6 – Field validation.....	82
5.4.8	Stage 7 – Conformity and safety assessment	83
5.4.9	Stage 8 – Authorization and entry into service	84
5.4.10	Stage 9 – Change and re-assessment.....	84
5.4.11	RACI summary tables	85
5.4.12	Milestone and evidence plan.....	85
5.4.13	Change and update governance	86
5.4.14	Exported constraints	87
6	Proposal for further Standardization	87
6.1	Guiding principles	87
6.2	Priority Activities prior to future standardization activities	87
6.3	Expected Benefits	88
7	Improvements & limitations of deliverable	89
8	Conclusions	91
	References	93

1 INTRODUCTION

Deliverable D6.4—*Proposal for the Certification Process of Perception Systems*—is produced within Europe's Rail Joint Undertaking (ERJU), Flagship Project 2 (FP2) **R2DATO**, Task/WP 6.4. It contributes to the overall objective of enabling Digital & Automated—up to Autonomous—Train Operation by defining a certification process enabling also AI--based perception (PER) on rolling stock. The work interfaces with adjacent activities, notably **WP 6.3 Architecture and System Specification** (foundational system/Perception architecture and requirements) and **WP 8.2 PER Safety Analysis** (hazard analysis and SIL needs), and aligns the proposed process with railway regulatory practice.

Background and motivation

PER functions are a prerequisite for higher Grades of Automation (GoA3/4), where trains must reliably detect and interpret their environment (e.g., obstacles, signals, and infrastructure conditions). Bringing such functions into service in European operations requires reconciling safety assurance, interoperability, and regulatory obligations. Compared with conventional deterministic systems, machinelearning (ML) components pose characteristic challenges data-dependence, non-determinism, lifecycle change through updates, and explainability—that must be addressed without departing from established rail certification principles. D6.4 responds to these challenges by consolidating safety and AI guidance into a process that can be scrutinized by authorities and assessment bodies.

Regulatory and normative context (high level)

The process is designed to fit the European rail conformity framework: **Technical Specifications for Interoperability (TSIs)**—especially LOC&PAS, CCS, and OPE—for subsystem/IC evaluation; the **Common Safety Method for Risk Evaluation and Assessment (CSMRA)** for risk acceptance; and forthcoming **EU AI Act** obligations for high-risk AI systems. The D6.4 deliverable notes the cautionary status of AI/ML in **EN 50716** and references cross-domain safety/AI sources applied throughout D6.4: **IEC 61508**, **ISO/IEC TR 5469:2024**, **DIN DKE SPEC 99002** (terminology), and **DIN DKE SPEC 99004** (rail ODD taxonomy). These sources inform claims allocation, lifecycle activities, and acceptance metrics.

Inputs and prior work

D6.4 synthesizes contributions from European and cross-sector initiatives to ground the workflow in state-of-the-art practice. **TAURO D1.2** informs the use of a systems-engineering **Vmodel** complemented by an MLspecific “W” (learning assurance), scenario abstraction (functional→logical→concrete), and a layered safetycase structure adapted from PEGASUS. **Confiance.ai** (Body of Knowledge) contributes method and tooling perspectives for robustness, explainability, data governance, and continuous monitoring. **DEEL** creates tools and processes for integrating machine learning (ML) techniques in safety-critical systems by addressing their high-level properties. **safe.trAI**n provides a realistic GoA3/4 driverless train example, including architectural patterns, quantitative acceptance discussion, and lessons for building evidence across simulation, HiL/SiL, and field testing. Where applicable, D6.4 aligns these sources with rail domain constraints and terminology.

Objectives of WP 6.4 / this deliverable

Task 6.4 aims to **develop a first fully traceable, evidence-based certification framework** for AI-based Perception system for railway in line with the defined scope of the grant agreement. Specifically, D6.4 delivers (i) a **general certification workflow** with clearly defined interfaces to ERA/NSAs/NoBos/DeBos/ASBOs/Stakeholders; (ii) an **extended, staged workflow (Stages 0–9)** that operationalizes activities from concept and ODD scoping through authorization and change management; and (iii) roles-specific responsibilities and artefacts (e.g., ODD definitions, datasets and labeling governance, HDmap quality and change control, simulation fidelity, performance & uncertainty reporting, and safety case evidence mapping). The deliverable explicitly **depends on WP 6.3** (requirements/architecture) and **WP 8.2** (safety requirements and SIL allocation) to maintain end-to-end traceability from hazards to evidence.

Process overview (summary)

The workflow is structured to mirror rail conformity steps while adding PER-specific assurance with optimal AI related extensions. Early stages establish governance, ODD baselines, and data foundations; middle stages cover safety/risk analysis and multimodal V&V (lab/simulation/field) with scenario-based coverage; later stages handle conformity and safety assessment, authorization/placing in service or placing on the market, and change control (including model/data updates). The process clarifies when and how simulation evidence is admissible, how acceptance metrics (e.g., detection/false alarm, latency, availability, uncertainty) map to safety goals, and how responsibilities are distributed among industrial stakeholders and assessment bodies.

Document structure

The remainder of this document is organized as follows: **Chapter 2** compiles the inputs used; **Chapter 3** states the goals of WP 6.4; **Chapter 4** summarizes the state of the art (IEC 61508, ISO/IEC TR 5469, DIN DKE SPEC 99002/99004, DEEL, TAURO D1.2, Confiance.ai, safe.trAIIn); **Chapter 5** details inputs into the process (regulatory/normative context, requirements, architecture, safety requirements); **Chapter 6** describes the certification process (actors, general workflow, PERspecific aspects, extended stages, RACI/milestones/change governance); **Chapter 7** proposes further standardization topics; **Chapter 8** discusses improvements and limitations; **Chapter 9** provides conclusions.

2 GOALS OF WP 6.4

The objectives of Task 6.4 are to Develop a fully traceable, evidence-based certification framework for AI-enabled perception functions in rail applications that can be scrutinized by the System Pillar and serve as a basis for European and international standardization.

Specific Objectives

- **Leverage existing knowledge** – Build on the Shift2Rail TAURO consolidation of external inputs to capture state-of-the-art certification requirements for perception algorithms.
- **Align with new operational needs** – Incorporate the use cases collected in WP5, ensuring the certification framework remains relevant across emerging operational design domains (ODDs).
- **Embed safety & hazard analysis** – Integrate the risk assessments produced in WP8, establishing a systematic link between identified hazards and corresponding assumptions, safety requirements and SRACS.
- **Define an end-to-end process** – Document a reproducible workflow (see Section 6) that spans requirement elicitation, architecture definition, quality assurance methods, system integration, system validation and safety assessment.
- **Enable adoption & standardization** – Deliver a reference process package by M36 that can be reviewed by the System Pillar and proposed for inclusion in relevant standards (EN 50716, IEC 61508, ISO/TR 5469).

3 SUMMARY OF THE STATE OF ART

3.1 IEC 61508-1:2010 FUNCTIONAL SAFETY

IEC 61508-1:2010 Functional safety of electrical/electronic/programmable electronic safety-related systems overview

IEC 61508 [7] is the international standard for the functional safety of electrical/electronic/programmable electronic safety-related systems (E/E/PE or E/E/PES). It provides a framework for designing systems that reduce risk to acceptable levels through safety functions and lifecycle management.

The standard introduces the concept of Safety Integrity Levels (SIL), which quantify the reliability of safety functions. SILs range from 1 (lowest) to 4 (highest) and are determined based on risk assessments and required risk reduction.

Demand Modes in IEC 61508

IEC 61508 distinguishes between **two modes of operation** for safety functions:

1. **Low Demand Mode**
2. **High Demand Mode / Continuous Mode** (similar to high demand in terms of metrics)

Regarding Low Demand Mode

Definition: The safety function is demanded less than once per year.

Typical Use: Emergency shutdown systems, fire suppression, airbag deployment.

Reliability Metric: Uses PFDavg – *Average Probability of Dangerous Failure on Demand*.

Risk Scenario: A hazardous event occurs while the safety system is in a failed state.

SIL Determination: Based on the probability that the system will fail when needed.

Regarding High Demand Mode

Definition: The safety function is demanded more than once per year, often daily or continuously.

Typical Use: Interlocking, process control systems...

Reliability Metric: Uses PFH – *Probability of Dangerous Failure per Hour*.

Risk Scenario: A failure directly leads to a hazardous event due to frequent operation.

SIL Determination: Based on the frequency of dangerous failures over time.

SIL Targets by Mode

SIL Level	Low Demand Mode (PFDavg)	High/Continuous Mode (PFH)
SIL 1	$\geq 10^{-2}$ to $< 10^{-1}$	$\geq 10^{-6}$ to $< 10^{-5}$ failures/hour
SIL 2	$\geq 10^{-3}$ to $< 10^{-2}$	$\geq 10^{-7}$ to $< 10^{-6}$ failures/hour
SIL 3	$\geq 10^{-4}$ to $< 10^{-3}$	$\geq 10^{-8}$ to $< 10^{-7}$ failures/hour
SIL 4	$\geq 10^{-5}$ to $< 10^{-4}$	$\geq 10^{-9}$ to $< 10^{-8}$ failures/hour

Key Considerations between high demand and low demand mode

Aspect	Low Demand Mode	High Demand / Continuous Mode
Diagnostic Requirements	No minimum diagnostic test rate; proof testing is key	Continuous diagnostics required; diagnostic coverage must be high (redundant architectures, watchdog timers, self checking mechanisms...)
Systematic Capability (SC)	Must meet safety requirements for the assigned SIL	Same safety requirements apply, but more emphasis on real-time fault detection
Design Complexity	Often simpler operationally, but proof testing and fault tolerance are critical	More complex due to real-time performance and frequent activation
Failure Scenario	Failure + demand must coincide	Failure alone can lead to hazard
Testing Requirements	Periodic proof testing to detect latent faults	Online diagnostics and frequent testing to ensure safe operation
Sector Usage	Common in process industries (e.g.,	Common in machinery, automotive including AI-based perception systems,

	emergency systems)	shutdown	Railways for Interlocking, Automatic Train Protection function
--	-----------------------	----------	---

In the context of Railway perception system used as “parachute” / emergency function, if the safety function is demanded less than once per year, low demand mode could be considered even if this is not usual within Railway industry (Low demand mode does not exist in EN50126/50129).

In the context of perception system used for the Use Cases defined in FP2-R2DATO D5.2, i.e. object detection, signal reading, surveillance of location associated with increased risk of collision, detection of infrastructure defects detection and anomaly detection, automatic joining of multiple units, “High Demand/Continuous Mode” could be considered which corresponds to the approach taken in the Safety analysis for Perception System, FP2-R2DATO D8.2.[18]

3.2 ISO TR 5469:2024

The Technical Report ISO/IEC 5469:2024 Artificial intelligence — Functional safety and AI systems provides recommendations and guidelines to address functional safety[8]. AI technology classes are identified:

- From class I: this is assigned if the AI technology can be developed and reviewed using existing functional safety methods and standards.
- To class III: this is assigned if the AI technology cannot be developed and reviewed using existing functional safety methods and standards.

The second criteria to classify the AI application is the usage level. This parameter can vary between usage level A1 to usage level D.

- Usage level A1 should be assigned when the AI technology is used in a safety relevant system and actively contributes to take automated decisions.
- Usage level D is assigned if the AI technology is not involved in safety relevant functions.

The combination of the two indexes is then used within a matrix to evaluate if the usage of a certain AI technology is adequate.

AI Technology Class = > AI application and usage level	AI technology Class I	AI technology Class II	AI technology Class III
Usage Level A1 ^a	Application of risk reduction concepts of existing functional safety International Standards possible	Appropriate set of requirements ^c	At the time of publication of this document no appropriate set of properties with related methods and techniques is known to achieve sufficiently reduction of risk
Usage Level A2 ^a		Appropriate set of requirements ^c	
Usage Level B1 ^a		Appropriate set of requirements ^c	
Usage Level B2 ^a		Appropriate set of requirements ^c	
Usage Level C ^a		Appropriate set of requirements ^c	
Usage Level D ^b	Application of risk reduction concepts of existing functional safety International Standards		
^a Static (offline) (during development) teaching or learning only.			
^b Dynamic (online) teaching or learning possible.			
^c The appropriate set of requirements for each usage level is established by application of risk reduction concepts of existing functional safety International Standards and additional considerations based on the literature review performed in Clauses 8, 9, 10 and 11. Examples are provided in Annex B.			

Figure 1: Example of AI classification table 1 cited from ISO TR 5469

For example, if on one hand the AI technology class I is considered adequate for a usage level A1, on the other hand the AI technology class III is recommended only for usage level C (as far as an appropriate set of requirements is established).

The Figure 2 hereafter shows visually the relationship between safety criticality and the possible level of autonomy of AI as described in the ISO TR 5469

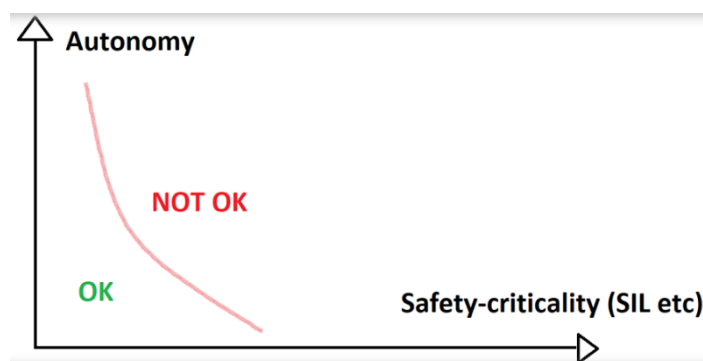


Figure 2: Relationship between “Autonomy” and Safety-criticality according to ISO TR 5469

A main concept in the report is deriving acceptance criteria for the three-stage realization principle as defined by the ISO/IEC 22989:2022 standard. The three stages are:

- Data acquisition (input tasks);
- Knowledge induction from data and human knowledge (learning tasks including training);
- Processing and generation of outputs (processing tasks including inference)

For each of the three stages it is important to define the desirable properties, detail the methods and techniques to be used, and to establish acceptance criteria.

ISO/IEC 5469:2024 details verification and validation techniques (Clause 9), control and mitigation measures (Clause 10), and processes and methodologies (Clause 11).

Clause 10 emphasizes the importance of robust AI system architecture for safety-critical applications. It highlights that architectural resilience is preferable to solely improving AI model quality and provides some guidance on subjects to be considered.

1. AI Subsystem Architectural Considerations

- ML redundancy with output voting or aggregation.
- Safe subset enforcement via supervisory constraints.
- Backup functions using non-AI techniques (e.g., failsafe actions).

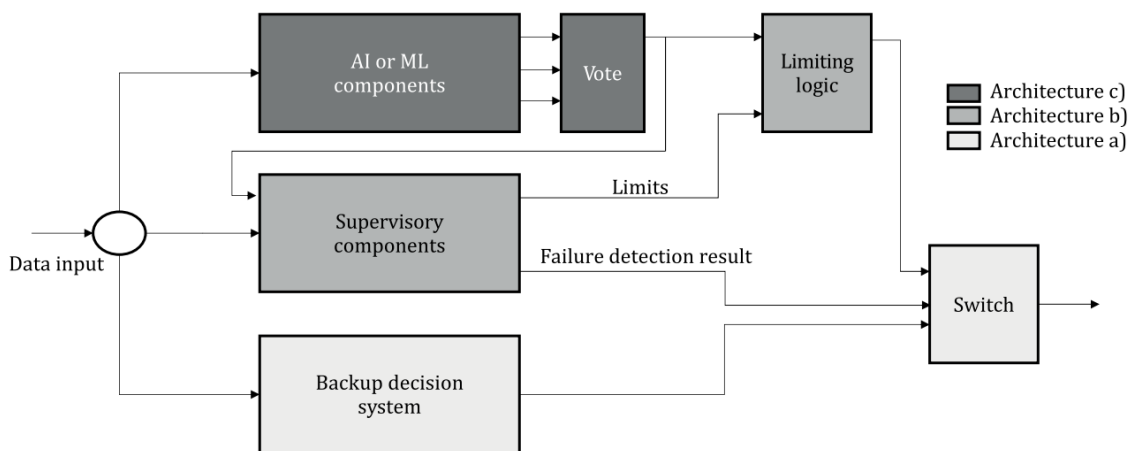


Figure 3: Architectural patterns for systems using AI technology components cited from ISO TR 5469, figure 5

2. Detection mechanisms

- Monitors detect unsafe AI behavior and trigger safe fallback.
- ODD input detection is critical due to generalization limits.
- Anomaly detection and adversarial robustness are key concerns.
- Use of runtime checks and input validation to detect attacks.

3. Supervision functions with constraints

- Constrains AI to operate within a safe envelope.
- Uses control theory methods like barrier functions.
- Supports fallback to stabilizing controls when needed.

4. Redundancy, ensemble concepts, and diversity of AI/ML components

- Types: structural, time, functional, and combined redundancy.
- Techniques: analytical redundancy, N-version programming, ensembles.
- Diversity strategies: different teams, data, hardware, sensing.
- Use of diversity metrics to select robust ensemble members.
- Simulation-based validation to ensure independence of models.

5. AI System design with statistical evaluation

- AI performance is statistically characterized due to high-dimensional inputs.
- Use of confidence intervals for error rates under defined conditions.
- Risk-based sampling for rare but high-impact events.
- Treat AI systems as probabilistic models for safety evaluation.

6. Increase the reliability of components containing AI technology

Identifies supporting technologies to enhance reliability of trained AI systems during deployment:

- robust learning
- optimization and compression technologies
- attention mechanisms
- protection of data and parameters

Next step:

The next step of standardization should replace this technical report with three technical specifications (under development):

- ISO/IEC AWI TS 22440-1 – Artificial intelligence — Functional safety and AI systems — Part 1: Requirements
- ISO/IEC AWI TS 22440-2 – Artificial intelligence — Functional safety and AI systems — Part 2: Guidance
- ISO/IEC AWI TS 25223 Information technology – Artificial intelligence – Guidance and requirements for uncertainty quantification in AI systems

3.3 DIN DKE SPEC 99002

The use of artificial intelligence in guided transportation systems, notably driverless trains, predictive maintenance, and service planning, is increasing. Thus, DIN DKE SPEC 99002 aims to establish uniform and clear terminology to facilitate effective communication across this cross-sectoral domain.

The DIN DKE SPEC 99002 provides standardized definitions for essential terms related to artificial intelligence (AI) in railway systems. Its primary goal is to establish a consistent and clear communication framework among manufacturers, suppliers, operators, certification bodies, and the scientific community. Rather than redefining terms anew, the specification compiles existing definitions from relevant standards and verifies their consistency and applicability within railway-specific contexts.

Key Definitions:

- **Artificial Intelligence (AI)** refers to the research and development mechanisms enabling machines to autonomously and adaptively make decisions.

Example (verbatim from standard 3.1 based on DIN EN ISO/IEC 22989:2024-09, modified): "Research and development of mechanisms and applications of AI systems."

- **The AI System** encompasses systems that exhibit varying degrees of autonomy and generate predictions, recommendations, or decisions based on their inputs.

Example (verbatim from standard 3.2 based on the EU AI Act): [5] "Machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."

Additional terms defined within the specification include those related to Machine Learning (ML), Deep Learning, data quality and management (such as Training Data, Validation Data, and Test Data), robustness of AI systems, safety approval processes, conformity assessments, redundancy measures, as well as railway-specific operational challenges and contexts like Driver Monitoring Systems (DMS), Collision Prediction, and Grades of Automation (GoA).

The annexes Annex A.1 until A.13 of the document contain comparative tables and overviews of various AI-related terms and definitions found in existing standardized documents, illustrating consistency checks and alignment across standards.

Significance of DIN DKE SPEC for this Document:

The document clearly highlights the necessity of consistent terminology definitions across various standards. Horizontal standardization of AI-related terminology across sectors such as automotive, aviation, or rail would be beneficial. If horizontal standardization proves infeasible, sector-specific standardization should be pursued, representing an essential task for European standardization. Given the requirement for consistent terminology in this document, and the intent not to create new definitions, the terms from DIN DKE SPEC 99002 are adopted and utilized.

3.4 DIN DKE SPEC 99004

The DIN SPEC 99004 titled "Specification of Operational Design Domain (ODD) in Rail" defines a comprehensive taxonomy for the Operational Design Domain relevant to various railway applications involving AI systems. The term Operational Design Domain refers to the explicit operating conditions under which an AI-based system is specifically developed and intended to operate.

In the railway context, the ODD is not only applicable to automated driving systems such as Automatic Train Operation (ATO) but also to other applications like visual inspection for automated monitoring of train maintenance. The purpose of DIN SPEC 99004 is to provide a standardized and sufficient taxonomy for describing operational conditions for AI applications in rail transport. The document does not include specific ODD descriptions for particular routes, nor does it specify detailed safety requirements for functionalities like the detection of persons in hazardous areas.

DIN SPEC 99004 primarily targets users, manufacturers, operators, testers, and developers involved with AI-based systems in rail applications. The document highlights the importance and benefits of a structured ODD taxonomy, which is essential for conducting risk assessments, supporting system development, scenario testing, and conformity assessments. A clear and explicit taxonomy supports the structured definition of ODDs, facilitating the development and validation of safety-critical systems.

Key aspects of the taxonomy include:

- Categorization of specific operational conditions and scenarios.
- Detailed identification of factors affecting the operation of automated and AI-based systems, covering infrastructure, environmental, and dynamic elements.
- Framework for supporting risk assessments concerning safety and security.
- Guidance for specifying hardware and software requirements based on operational conditions.

The document promotes the reusability and comparability of ODD descriptions by providing a fundamental taxonomy applicable to multiple use cases within rail operations. This standardization fosters a common understanding across different applications.

Chapter Structure of DIN SPEC 99004:

1. Scope
2. Normative references
3. Terms and definitions
4. Symbols and abbreviations
5. Introduction to ODD

This chapter provides an overview of the Operational Design Domain (ODD), highlighting its conceptual foundations, significance, and practical implications within rail applications. It explains how an explicit definition of operating conditions can support AI-driven systems and addresses the essential role of ODD in safety, security, and compliance processes. The chapter outlines the primary objectives and uses of ODD definitions, setting the stage for detailed lifecycle discussions.

6. ODD Lifecycle

The ODD lifecycle chapter details a structured process for defining, refining, and managing ODD throughout various phases of system development and operation. Initially, it addresses general lifecycle management tasks including configuration management, tracking, and traceability. The chapter subsequently describes specific phases such as Operational Concept, emphasizing establishing clear project goals and context. System Definition and Operational Context phases focus on defining detailed scenarios and conditions under which the AI system operates. Risk Analysis and Evaluation assess potential safety, security, and ethical risks linked to the defined ODD, influencing subsequent specifications. The Specification of System Requirements phase translates these insights into explicit technical, operational, and safety requirements aligned with the ODD. ODD Architecture involves identifying critical attributes influencing system operation and systematically reducing complexity. Hardware and Software Requirements are then explicitly derived from these attributes, ensuring effective data collection, verification, and operational readiness. The System Design and Implementation phase addresses data collection and testing strategies to guarantee ODD's validity. System Validation evaluates compliance through targeted testing within

defined operational scenarios. Deployment phase details safe, structured system integration into real-world environments, including operational and maintenance considerations. Lastly, the Operation and Monitoring phase ensures continuous real-time alignment of system performance with defined ODD parameters, enabling prompt responses to evolving conditions.

7. Use of ODD Definition Format

This chapter discusses the importance of using a standardized format for ODD definitions to facilitate consistency, clarity, and interoperability among various stakeholders. It outlines the benefits of human and machine readability, compatibility with existing standards, and the necessity for a common format to support efficient data exchange, scenario testing, and system integration.

8. Taxonomy for ODD

The taxonomy chapter provides a detailed framework for categorizing and describing operational conditions relevant to railway applications. It introduces three primary groups: scenery elements, environmental conditions, and dynamic elements. Scenery elements include stationary features such as tracks, signals, platforms, junctions, and infrastructure components like bridges and tunnels. Environmental conditions encompass a broad range of factors, including weather parameters (temperature, rainfall, snowfall, wind), illumination conditions (daylight, artificial lighting), and atmospheric particulates (fog, dust, pollution). Dynamic elements capture both predictable and unpredictable moving entities such as trains, pedestrians, vehicles, animals, and temporary obstacles (e.g., construction equipment). Each category is detailed with specific attributes, enabling precise specification and differentiation of operational scenarios. The comprehensive hierarchical structure of this taxonomy supports rigorous and repeatable scenario definitions, enhancing clarity and facilitating accurate communication across diverse stakeholders involved in system development and validation.

9. Exemplary Use Cases

This chapter illustrates practical applications of the ODD taxonomy through concrete examples, such as obstacle detection for automated driving systems, visual inspection for automated train maintenance, and remote train operation systems. Each example demonstrates how taxonomy can be practically applied to define operating conditions, manage system development, support validation processes, and achieve compliance with safety standards. For instance, the obstacle detection system example shows how environmental and dynamic elements are systematically analyzed to determine potential hazards. The visual inspection use case demonstrates the application of scenery and environmental attributes to ensure thorough inspection. The remote driving scenario highlights the taxonomy's utility in specifying operational limits and necessary technological requirements for safe remote train operation.

10. Annexes and Bibliography

Annexes provide supplementary information and guidance on specific aspects of ODD definitions, including additional technical details, best practices, and extended examples.

The bibliography lists relevant sources and standards referenced throughout the document, offering resources for further exploration and research.

3.5 DEEL PROJECT

DEEL[1] is [1]a program gathering worldwide 27 partners, including academic researchers, data scientists and thematic engineers specialized in operational safety and certification.

The objectives of DEEL are three-fold:[1]

1. To create internationally recognized scientific and technological knowledge on the explainability and robustness of artificial intelligence algorithms in an automatic and statistical learning context.
2. To create conditions for the rapid transfer of work to its industrial partners through the development of skills and the provision of tools (available in a github).
3. To make the scientific community and the general public aware of the emblematic scientific challenges built with the partners.

DEEL aims at integrating machine learning (ML) techniques in safety-critical systems across automotive, railway, and aerospace industries. ML has transformed problem-solving capabilities in previously challenging technical areas like natural language processing and pattern recognition in large data volumes. However, its inclusion in certified systems presents unique challenges.

3.5.1 Challenges of ML in Certified Systems

Seven main challenges of ML certification, named as **High-Level Properties**, are addressed:

1. **Probabilistic Evaluation:** It is crucial to extrapolate statistical performance of ML models into operational safety indicators, considering real-world probabilities—like ensuring the low likelihood of catastrophic events in railway systems.
2. **Resilience:** Ensuring safe operations in case of system failures is vital, necessitating architectures that allow ML error recovery and monitor unexpected or erroneous inputs.
3. **Specifiability:** Defining expected system behaviors using learning quality datasets to specify functions that the system must fulfill is essential.
4. **Quality and Representativeness of Data:** Establishing clear criteria to ensure data reflects necessary operational environment variations for ML training, validation, and testing.
5. **Explicability:** Generating explanations to assist human operators in tracking and making informed decisions, especially under system failure conditions.
6. **Robustness:** Evaluating how ML systems withstand perturbations or anomalous inputs ensures coherent responses and protection against classification errors caused by adversarial examples.
7. **Verifiability:** Implementing reliable methods for validating that ML systems correctly perform intended functions and meet established safety requirements.

3.5.2 Certification practices in the railway industry

In accordance with European regulations, the railway system is broken down into structural subsystems and functional subsystems. The standards given in **Fehler! Verweisquelle konnte nicht gefunden werden.** are applicable to the hardware and software parts of systems/sub-systems/equipment.

Table 1: simplified requirements for railway certification

	Interoperability Directive (EU) 2016/797	Safety Directive (EU) 2016/798	
System	Set of TSI (Technical Specification for Interoperability)	CSMs (Common Safety Methods 402/2013/EC) NNSR/TR (Notified National Safety Rules /Technical Rules)	Law
HW/SW	Standards called by TSI	50126, 50128/50657, 50129	Rules, Methods, Standards
Mechanical	Standards called by TSI	EN 15566, EN 15551, EN 16286-1, EN 15085-5, ...	
Fire	Standards called by TSI	EN 45545	
Etc.			

Before an application for Safety approval can be considered, an independent safety assessment of the system/sub-system/equipment and its Safety Case shall be carried out to provide additional assurance that the necessary level of safety has been achieved. The resulting Safety Assessment Report must explain the activities carried out by the safety assessor to determine how the system/sub-system/equipment (hardware and software) has been designed to meet its requirements. It can also specify some additional conditions for the operation of the system/sub-system/equipment. In addition to the strict application of standards, European regulation 402/2013/EC requires railway operators to implement a “common safety method relating to the assessment and evaluation of risks” to all modification, whether technical, organizational or operational, of the railway system. The Common Safety Method describes how the safety levels, the achievement of safety targets and compliance with other safety requirements should be fulfilled.[6]

The risk acceptability of the system under assessment shall be evaluated by using one or more of the following risk acceptance principles:

1. Compliance with a recognized and context-specific code of practice (see Table 3) to deal with conventional risks that are not necessary to explicitly re-identify if there is no technical or functional innovation or evolution of the environment and context.
2. Comparison with a “similar reference system”. This principle existed previously to the CSM in France as GAME (“*Globalement Au Moins Equivalent*”): it is possible to accept a new system if it is shown that the deviations from the existing system, having the same functionalities and operating in a similar environment to the system to which it is compared, do not introduce new risks or increase existing risks.
3. Explicit risk analysis to demonstrate quantitatively or qualitatively that the risk is kept low enough to be acceptable. In the end, the Certification Authority is based on the opinion of an independent expert to evaluate in order to obtain the confidence.

3.5.3 Towards new certification practices in the railway sector

The CENELEC EN-50128 standard recognizes that default prediction (trend calculation), defect correction, and maintenance, and monitoring actions are supported very effectively by Artificial Intelligence-based systems in various parts of a system. This does apply exclusively to SIL0 (Safety Integrity Level) as it is not recommended for higher SIL levels. The regulatory means to introduce AI in EU railway sector still needs to be defined. This is illustrated by the latest Locomotive & Passenger train TSI (LOC&PAS TSI), which deals with “innovative solutions”. Indeed, to allow technological progress, it may be necessary to introduce innovative solutions not complying with defined specifications (i.e. that do not meet TSI specifications or to which TSI evaluation methods cannot be applied). In this case, new specifications and / or new evaluation methods associated with these innovative solutions must be developed and proposed. Innovative solutions can involve the “Infrastructure” and “Rolling Equipment” subsystems, their parts and their interoperability components.

3.5.4 Towards certification of ML

The following approaches for ML certification are proposed by DEEL:

- **Process-based certification:** the completion of a predefined development assurance process during the development of a product is the assurance that this product complies with the requirements laid down in the certification basis. Most of the standards currently used for certification are “process-based”. In this approach, High-Level Properties (HLPs) are demonstrated as part of the full process.
- **Property-based certification:** the demonstration that a predefined set of properties is met by a product is the assurance that this product complies with the requirements laid down in the certification basis. In this approach, the HLPs may be some (or all) of the properties to be demonstrated. Even though there is a growing interest for this approach, one of the difficulties is to prove that the selected set of properties completely covers the desired certification objectives. A clear consensus on an acceptable set of properties has not been reached yet, but “property-based” certification is in line with the Overarching Properties initiative, and it remains an option considered for the future.

Those two approaches are not exclusive and can be combined. Whatever the certification framework chosen, we believe that the ability to demonstrate the HLPs will be a key enabler for

ML-based system certification. Depending on the system and the certification strategy, the HLPs to be demonstrated, as well as the depth of demonstration may vary, but in any case, this list of HLPs summarizes the main challenges of ML certification.

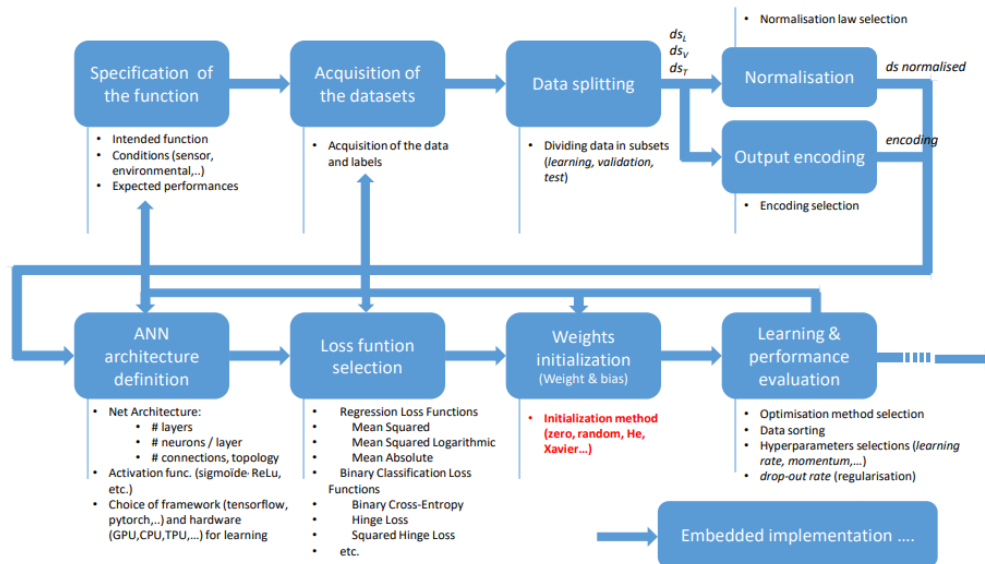


Figure 4: Simplified machine learning workflow (from DEEL white paper)

3.5.5 Opportunities for ML in industry

Significant opportunities exist for applying ML in automotive, railway, and aerospace sectors, such as enhancing predictive maintenance, quality control, autonomous driving (ATO up to GoA4), and optimizing transport network operations. However, incorporating ML in safety-critical systems demands evolving certification practices to address these challenges.

3.5.6 DEEL way-forward

DEEL project tackles certification challenges in ML-based systems for safety-critical industries. Providing explanations about ML models is a key feature for a better mastering of the systems using them. When these systems become critical, focus must be set on the explanations allowing a human operator to supervise the decisions taken by the system. Main challenges with the certification of the ML system are about interpretability of the communicated information and completeness of the communicated information.

Three types of analyses are proposed by DEEL: a detailed analysis of a typical ML-system development process, a similarities analysis between ML techniques and techniques already implemented in certified systems, and a backward analysis to point out ML-techniques and applications that do not show some of the other techniques' challenges.

Identifying high-level properties (HLP) for ML could lay the foundation for future certification methodologies, helping validation within existing regulatory frameworks. Collaboration across disciplines is essential for advancing ML system certification, initiating important research avenues.

The DEEL continues to tackle part of these HLP challenges, either from a scientific point of view, or following a bottom-up approach towards industrial use cases (e.g. railway signals images classification) within its certification workgroup.

Links of interest can be found in [1], [2] and [3].

3.6 TAURO D1.2 – PROPOSAL FOR CERTIFICATION METHODOLOGY OF SAFE FUNCTIONS BASED ON AI

3.6.1 Introduction and Context

The **TAURO D1.2** report, “Proposal for certification methodology of safe functions based on artificial sense,”[16] was prepared as part of the European Shift-2-Rail **TAURO** project. The report addresses how to certify **perception functions** based on artificial intelligence (AI) so that they can be safely deployed on railway rolling stock. With increasing levels of automation (up to GoA 4), trains need to interpret their environment – detecting track occupancy, obstacles, signals, and passenger behavior – in ways that were previously handled by human drivers. To achieve this, rail systems will employ **computer vision and other sensor-fusion algorithms** based on machine learning. These functions are safety critical: erroneous object detection could trigger false braking or missed obstacles, jeopardizing passengers and infrastructure.

However, AI driven perception has inherent challenges. Machine learning models act as “black boxes” whose decision logic may be opaque. Their behavior depends on data quality and can change over time as models are retrained. There are regulatory uncertainties in adapting existing functional safety frameworks to AI. The **EU Artificial Intelligence Act** treats such perception systems as **high-risk AI**, requiring specific obligations for risk management, data governance, transparency and human oversight. Traditional railway standards (TSIs, CSMRA[6] and CENELEC EN 50126/50128/50129) offer robust processes for deterministic systems but lack guidance on non-deterministic AI. **TAURO D1.2** therefore proposes a harmonized certification methodology combining classical safety engineering with AI-specific assessment. It draws on automotive experience (ISO 26262, SOTIF), the **PEGASUS** project for automated driving, the **EASA** AI concept paper, and current research on trustworthy AI, aiming to tailor these concepts to rail.

3.6.2 Normative and Regulatory Framework

The report reviews the evolving normative landscape that certification must navigate. Classical **functional safety** is governed by IEC 61508 [7] and its rail adaptation **EN 50126/50128/50129**, which prescribe RAMS processes and Safety Integrity Levels (SIL) for control systems. For AI perception, **ISO 26262** (automotive functional safety) and **SOTIF** (ISO 21448) are informative: ISO 26262 addresses hazards due to system malfunctions, whereas SOTIF covers hazards arising from correct behavior under unconsidered conditions or limitations of environmental perception. The **EU Artificial Intelligence Act** employs a risk-based classification and requires high-risk AI to undergo conformity assessment, maintain technical documentation, ensure data quality and enable human oversight. The report also notes upcoming standards such as **ISO/IEC TR 5469** on functional safety of AI systems, which align AI lifecycles (data collection, learning, inference) with functional safety and emphasize verification, validation and fallback architecture.

Importantly, the report identifies that the rail domain has limited resources for bespoke AI standards and should therefore collaborate horizontally with other sectors. Harmonized **horizontal standards** (applicable to multiple industries) can define general AI development processes, data management practices and trustworthiness criteria, which can then be tailored to rail via parameterization. Aligning with automotive and aerospace activities (e.g., EASA's AI roadmap) ensures technology convergence and leverages existing tools and methods.

3.6.3 Trustworthy AI: Principles and Dimensions

A core premise of TAURO D1.2 is that certification requires **trustworthy AI**. Classical performance metrics (e.g., accuracy) are insufficient; AI systems must also satisfy legal, ethical and social requirements. The report adopts the European Commission's **Ethics Guidelines for Trustworthy AI**, which define three overarching components: the AI system must be **legal** (complying with all applicable laws), **ethical** (adhering to principles such as non-maleficence and justice), and **robust** both technically and socially. These principles translate into several dimensions that need to be addressed in development and certification:

- **Accuracy and Robustness:** AI must perform reliably across diverse contexts and be resilient to adversarial examples or distributional shifts. Test criteria should define target performance ranges and margin of error.
- **Privacy:** Data used for training and operation must be handled under strict privacy-preserving policies, ensuring compliance with GDPR and preventing sensitive information leakage.
- **Transparency and Explainability:** Users and auditors must be able to understand the system's behavior. This includes the concepts of *interpretability* (human comprehensible reasoning) and *completeness* (the explanation covers all relevant factors). The degree of required transparency may differ by stakeholder: developers need model-internal explanations, while operators may require high level rationale and confidence scores.
- **Accountability and Reproducibility:** The AI development process must produce audit trails enabling responsibility assignment. Results should be repeatable across runs.
- **Fairness and Ethics:** The system must minimize unintended bias; this demands careful analysis of data for discriminatory patterns and balanced representation.
- **Specification:** AI models must implement the intended function correctly; *specification errors* occur when the model optimizes an unintended objective due to ambiguous requirements.
- **Assurance and Monitoring:** Runtime monitors or supervisors should detect failures or anomalous behavior and trigger safe states (e.g., emergency braking). These are critical for high-risk AI where perfect prediction cannot be guaranteed.

The report distinguishes between **intrinsic trust** – gained through transparency and explanation of the model's inner workings – and **extrinsic trust** – gained through empirical validation and performance evidence. Certification should incorporate both: explanatory tools and model introspection build intrinsic trust for designers and auditors, while extensive testing, KPIs and statistical metrics establish extrinsic trust for operators and regulators.

3.6.4 Development Methodology: V-Model, Scenario Abstraction and Safety

3.6.4.1 V-Model as Structural Backbone

The **V-Model** from systems engineering and safety-critical software development provides the structural backbone for the proposed methodology. The left arm of the V addresses **requirements analysis and system decomposition**, while the right arm covers **integration, verification and validation**. TAURO D1.2 adapts this model for AI enabled functions. The process begins with collecting use case requirements, including operational design domain (ODD) parameters, performance targets, safety requirements (SIL/TFFR) and legal/ethical constraints. These requirements inform the design of the perception function, its architecture and interfaces. Systems are decomposed hierarchically: at vehicle level, high-level functions are defined and then allocated to subsystems (sensors, computing unit, communication interfaces). This ensures traceability of safety requirements down to individual components and allows early identification of hazards and misuse cases.

3.6.4.2 Scenario Abstraction and Data Processing

A key innovation is the **scenario-based approach** for test planning. The methodology distinguishes three abstraction levels:

1. **Functional Scenarios:** High-level descriptions in natural language, capturing the context and actors (e.g., “train approaching a level crossing with a road vehicle waiting”).
2. **Logical Scenarios:** Parametric representations describing the ranges of relevant variables (e.g., speed 80–100 km/h, weather conditions, object sizes) and constraints. This level supports systematic test case generation.
3. **Concrete Scenarios:** Specific instantiations with fixed parameter values, serving as the input for simulation or field tests.

This hierarchical decomposition enables systematic test coverage: functional scenarios are derived from safety goals; logical scenarios define parameter spaces; concrete scenarios drive simulations. **Data processing** involves acquiring or synthesizing data to fill scenario spaces. If real data is insufficient, synthetic data generation (e.g., via simulators) is used to augment coverage. All data and scenarios are stored in a **database** that supports queries and ensures the traceability between requirements, test cases and results. This database also records metadata, labels, model configurations and test outcomes, forming the basis for auditability and later updates.

3.6.5 PEGASUS Safety Argumentation and Layered Structure

TAURO D1.2 adopts and adapts the **PEGASUS** project’s safety argumentation framework, originally developed for automated driving. The framework decomposes a top-level **Safety Goal** into testable elements and connects them to evidence (tests, simulations, analyses) via an argumentation structure. UseCases trigger requirement extraction; these requirements guide data processing, test derivation and evaluation. Evidence from tests validates the safety argument.

The safety case is organized in **five layers**:

- **Layer 0 – Acceptance Model:** Defines what “safe” means by specifying the ODD, applicable laws and standards, and expected nominal behavior. It also defines conditions requiring fallback modes.
- **Layer 1 – Top-Level Safety Goals:** High-level objectives derived from the acceptance model (e.g., “The perception system shall detect obstacles in its ODD with probability $\geq P$ within latency $\leq T$ ”).
- **Layer 2 – Logical Structure:** Decomposes safety goals into sub-goals and requirements for individual components. It systematically analyses cross-cutting dependencies (e.g., sensor coverage vs. processing capacity) and guides allocation to subsystems.
- **Layer 3 – Actions, Methods and Tools:** Specifies which design principles, verification methods, tools and mitigation measures are employed to fulfil the requirements (e.g., sensor redundancy, model training techniques, simulation platforms).
- **Layer 4 – Evidence:** Collects objective evidence, including test results, verification artefacts, hazard logs, audit trails, and formal proofs. Evidence is tied back to the relevant goals and used to argue compliance.

This layered argumentation provides a structured certification narrative. It makes explicit the assumptions and limitations of AI functions and ensures that each claim is supported by traceable evidence.

3.6.6 Data Management and Traceability

Because AI models learn from data, certification must ensure that the **data lifecycle** is controlled. TAURO D1.2 categorizes data into several types (raw sensor data, metadata, labels, trained models, scenario descriptions, virtual assets, test reports) and defines traceability requirements. Key aspects include:

- **Data Provenance:** Every data item must carry metadata about its origin, conditions, transformations and intended use. Graph databases may be used to record relationships between data sets, conversion steps and model versions.
- **Data Governance:** Data should be accessible and secure across organizational boundaries, avoiding silos. Governance processes must ensure integrity, availability and confidentiality.
- **Semantic Linking:** Data should be enriched with ontologies that define concepts (objects, terrains, environmental conditions) and their relations, enabling consistent interpretation across tools and teams.
- **Standardized Interfaces:** Wherever possible, data formats should adhere to international standards or de-facto conventions, reducing conversion overhead and ensuring interoperability.

The report stresses **data quality** as crucial for trustworthy AI. It identifies **measurement errors** (e.g., sensor noise, labelling mistakes) and **representation errors** (lack of representativeness across the population or ODD) as major risks. Datasets used for AI must be assessed for completeness, accuracy, consistency, timeliness and bias. Reliability (low variance) and validity (measuring the intended phenomenon) are essential. Dataset descriptions should accompany each dataset, detailing composition, collection procedures, preprocessing and limitations. The report notes that no

universal standard for dataset descriptions exists; establishing one would support certification by enabling transparent evaluation.

3.6.7 Test Environments and Validation Methods

TAURO D1.2 proposes a **multimodal test strategy** combining virtual, HiL (Hardware in the Loop), SiL (Software in the loop) and field tests. Each modality addresses different phases of the development life cycle and different maturity levels (TRL levels). To include both aspects of HiL and SiL also the term XiL is used.

3.6.7.1 Virtual Testing

Virtual testing involves simulating the environment, sensors and vehicle dynamics. Simulation models may use different data formats; a co-simulation platform coordinates interactions. Simulation capabilities vary across **sensor fidelity**, **environmental realism** and **vehicle dynamics**—with fidelity quantified as low, medium or high. Vehicle dynamics tools simulate train maneuvers and interactions with track geometry, while environment tools add infrastructure, weather and traffic conditions. Traffic simulation tools generate other actors (cars, pedestrians), controlled by behavioral models. A **test manager** orchestrates scenario execution and collects outputs.

For rail, virtual scenarios can be generated through three approaches: (1) re-using 3D assets from gaming engines (e.g., Train Simulator), which offers limited perspective and ODD; (2) creating bespoke scenarios with general 3D software (e.g., Blender), allowing flexible camera placement; or (3) adapting automotive simulators (e.g., CARLA) to the rail domain, providing built-in support for physics and sensors. Research projects (e.g., VALU3S) are exploring the use of Train Simulator for verifying light-signal detectors, demonstrating how diverse conditions (weather, time of day, occlusions) can be created and used to test model robustness.

3.6.7.2 XiL Testing

XiL testing integrates real components into a simulated environment. In **Software in the Loop (SiL)**, perception algorithms run on a computing platform and process simulated sensor streams. In **Hardware in the Loop (HiL)**, sensors or ECUs (Electronic Control Unit) are physically connected and receive input signals via domain-specific cabling. XiL setups can be **open-loop** (data flows in one direction; outputs are recorded) or **closed-loop** (model outputs feed back into the simulation, affecting subsequent inputs). Realtime capability is critical; data flows must mimic the latency and bandwidth of onboard systems. Test infrastructure must allow switching between sensor configurations, injection of faults, and monitoring of outputs.

3.6.7.3 Field and Proving Ground Testing

Field testing provides the ultimate validation stage, assessing the function under real operating conditions. The report distinguishes **proving grounds**—controlled environments where scenarios are scripted—and **open world field tests**, where the system is exposed to uncontrolled traffic and environmental conditions. Field testing assesses whether the results from previous test stages hold under real noise and complexity. Challenges include lack of controllability (interaction with unpredictable actors), lack of repeatability, limited scalability (need for digital maps and V2X infrastructure), and high cost. Field tests should be conducted only after sufficient maturity is achieved through virtual and XiL testing (i.e., after reaching adequate Technology Readiness Levels).

The report outlines base requirements for open-infrastructure tests: coverage of the scenarios defined in the ODD, compliance with regulatory and ethical frameworks, ability to collect operational data, and readiness to revert to safe states. Data from field tests feed back into the scenario database, enabling continuous improvement.

3.6.8 Evidence and Acceptance Criteria

Certification requires defining **acceptance criteria**. The report highlights the **Operational Design Domain (ODD)** as a central concept: it describes the set of conditions (infrastructure type, train type, permissible weather, traffic conditions) for which the AI function is designed. Evidence must show that the AI meets performance targets within its ODD. For each scenario, performance is evaluated through Key Performance Indicators (KPIs) such as detection rate, false alarm rate, latency and robustness under sensor faults. Safety requirements allocated to the AI software (e.g., probability of missed detection below a threshold) must be demonstrated. Evidence also includes hazard logs, traceability matrices, Safety Integrity Level targets and test coverage reports. Where possible, tests should move into simulation to reduce costs and enable large coverage. Nonetheless, field demonstrations remain necessary.

3.6.8.1 Procedure Pipeline: Trustworthiness Analysis, Learning Assurance, Explainability and Risk Mitigation

TAURO D1.2 extends the traditional V-model with a **procedure pipeline** structured into four pillars, reflecting **EASA's AI roadmap**. These are **trustworthiness analysis**, **learning assurance**, **AI explainability** and **AI safety risk mitigation (SRM)**.

Trustworthiness Analysis

This first pillar establishes the interface between ethical guidelines (e.g., EU AI Act[20]) and technical assurance. It begins with a **system definition** detailing:

1. **End-users and stakeholders**: identification of those interacting with the AI (train drivers, remote operators, passengers, maintenance staff).
2. **System architecture**: decomposition into subsystems (sensors, perception algorithms, HMI) and delineation of responsibilities.
3. **Operational Concept and ODD**: description of the intended operation, including geographical scope, environmental conditions, train types and limitations.
4. **Functional Analysis**: breakdown of high-level functions into sub-functions and allocation to subsystems and AI constituents.
5. **System Safety Assessment**: application of CSMRA and EN 50126 principles. Novel steps include assigning assurance levels (e.g., SIL, TFFR) to AI items; defining performance and reliability metrics; [6]analyzing behavior when inputs fall outside the ODD; and performing failure mode effects analysis (FMEA) on AI constituents. The analysis must also address **cybersecurity** (e.g., data poisoning) and **ethics** (preventing discriminatory outcomes). Evidence should show that training and test datasets reflect demographic distribution and potential future variations.

Learning Assurance (W-Model)

While the V-model covers the overall system, AI software development follows a **W-shaped model**: after system requirements are allocated to AI constituents, the AI development process branches into a separate loop. **Learning assurance** covers:

1. **Requirement Management**: Capturing operational design domain, performance, scalability, reliability and resilience requirements specific to AI. The AI must know how to react when ODD boundaries are exceeded (e.g., revert to failsafe mode or request intervention). Non-functional requirements, such as minimum detection probability, acceptable inference latency and memory constraints, must be specified. Regulatory authorities should clarify nominal operating conditions and exceptional scenarios (e.g., fires, floods) where AI is not expected to operate.
2. **Data Management**: Careful planning of data collection, labelling, preparation and partitioning into training, validation and test sets. For highly critical applications, training and test sets should be sampled independently to prevent information leakage. Data representativeness and completeness must be ensured. Data quality management processes from avionics (e.g., DO-200B) or EN 50657 may serve as models.
3. **Learning Process Management**: Preparation of training includes selecting model families (e.g., CNNs Convolutional Neural Networks, transformers), algorithms, bias/variance metrics, training environment (hardware and software) and expected performance. Generalization guarantees should be quantified; however, current theoretical bounds often require enormous data volumes, indicating a limitation that further research must address. During training and optimization, changes affecting model behavior must be documented. Bias–variance tradeoffs should be evaluated using resampling methods.
4. **Verification and Validation**: The trained model is evaluated on test datasets to verify performance and robustness. Stability of algorithms and sensitivity to parameter variations are assessed. Only when acceptable generalization is demonstrated can the model be integrated back into the system. The W-model thus adds a **learning process verification phase** between the traditional V-model verification stages, ensuring that AI development does not bypass established safety gates.

AI Explainability

Explainability ensures that human stakeholders can understand AI decisions. TAURO D1.2 notes that explainability requirements vary by role:

- **Certification authorities and safety investigators** need insights into the internal workings to diagnose failures. Techniques include filter visualizations, generative methods, maximally activating inputs, local approximations, activation visualization and saliency maps. These methods will evolve with research.
- **Operators** need high-level reasoning, confidence metrics and system status to intervene appropriately.
- **Developers** need granular explanations to debug and improve models.

The report emphasizes that explainability builds on learning assurance: without confidence in model training, explanations may be misleading. It also stresses **data recording**: operational data should

be logged to support post hoc analysis and continuous assurance. Logging must be designed to respect privacy and security constraints.

AI Safety Risk Mitigation (SRM)

SRM acknowledges that the AI “black box” may not always be fully explainable. Residual risks—due to incomplete transparency or model limitations—must be mitigated through **additional safety measures**. In the rail context, SRM may involve:

- **Fallback functions:** Classical deterministic functions or simpler algorithms that run in parallel and can take over when AI fails (e.g., conventional obstacle detectors as backup).
- **Supervisory functions:** Runtime monitors that compare AI outputs with expected ranges; if discrepancies occur, they trigger alarms or safe states.

The EASA paper notes that guidelines are currently tailored for **Level 1 AI applications** (human assistance) and that refinements will be needed to support higher autonomy. TAURO D1.2 recognizes the need to accumulate operational experience—initially by deploying AI as driver assistance and gradually expanding autonomy while collecting data to inform future standards.

3.6.9 Test Data Management and Harmonisation

Section 5.5 of TAURO D1.2 addresses **test data management**. For a coherent certification process, data from various test modalities (virtual, XiL, proving ground, field) must be harmonized. Challenges include different execution mechanics, varied measurement units, and inconsistent parameter definitions. A common **evaluation stage** should ingest results, compute KPIs, and determine pass/fail criteria. If coverage is insufficient, the process should feed back to scenario selection and additional testing. Harmonization requires objective definitions of test specifications, parameters and measurements, and mapping between terms across modalities. Future work must develop common data models and analysis tools capable of cross-modal analytics.

Identification of New Elements Requiring Updated Certification Process

Chapter 7 highlights elements unique to AI that require updates to conventional certification processes:

- **Trustworthiness and Model Assessment:** AI models must provide evidence of trustworthiness through auditability and accountability. This includes demonstrable adherence to ethical guidelines, traceable development and runtime behavior, and mechanisms to handle novel conditions.
- **Machine Learning Processes:** The AI development pipeline must be documented; automatic recording of training events (hyperparameters, dataset versions) should enable reproducibility.
- **Data Quality and Security:** Data quality must be assured; data must be protected against tampering and unauthorized access. This is particularly critical because adversarial data poisoning can compromise model behavior.
- **Bias and Fairness:** Models must be checked against potential discrimination (e.g., based on geography, age or race). Certified datasets should be diverse; measurement bias (single sensor type) must be mitigated by including multi-sensor data.

- **Continuous Change and Currency:** Environments evolve. Datasets must be updated to reflect new societal developments; outdated data should be removed. AI models must be regularly retested and, if necessary, retrained on updated certified data. Certification should require periodic review to ensure models remain current.
- **Qualification of Users:** Operators must be properly trained to avoid automation bias (overreliance on AI). Certification should specify the required qualifications and responsibilities for people working with AI systems.

These elements underline that certification is not a one-time event but an ongoing process encompassing the entire life cycle of AI systems.

3.6.10 Recommendations and Future Steps

TAURO D1.2 concludes with recommendations for standardization and future work. It emphasizes **horizontal standardization**, suggesting that, because railway budgets for research and standards are limited, the sector should participate in cross-domain standardization efforts. Aligning with automotive and aerospace activities can yield shared **development guidelines** and **quality assured processes** to tackle AI challenges such as data distribution shifts, incomprehensible behavior, rare critical situations, unreliable confidence information, brittleness of deep neural networks, inadequate separation of test and training data, dependence on labelling quality and insufficient safety oriented metrics. The report proposes that many of these guidelines are already partly addressed in TAURO and could be consolidated into an international or European standard. They could be integrated into **ISO/IEC 5469** or European standards specifying trustworthiness requirements under the AI Act.[20]

A second recommendation is to define a **modular Operational Design Domain**. Based on the **ASAM ODD** approach used in automotive, the ODD should be modularized into elements such as environment, scenery, human actors (poses, trajectories), etc. Standardizing ODD descriptions facilitates scenario generation, comparison and reuse of machine learning components. CEN/CENELEC JTC 21 WG4 has initiated work on ODD standardization; the railway domain should contribute, for example, by defining a **driverless regional train ODD**. The R2DATO project could prepare a pre-standard for ODD definitions and scenario descriptions (possibly via CEN/CLC TC9X SG34 SG2). This would support the reuse of perception components across different rail applications and enable interoperability.

Finally, the report acknowledges that its methodology is an initial proposal and not yet sufficient for complete certification of AI systems. More research is required to refine generalization guarantees, develop robust explainability tools, define common datasets and scenario repositories, and specify performance criteria for new rail use cases. Accumulating operational experience by deploying AI as assistance functions and progressively increasing autonomy is proposed as a path to building evidence for future certification.

3.6.11 Conclusion

TAURO D1.2 provides a comprehensive proposal for certifying **AI-based perception functions** in the railway domain. It integrates classical safety processes with AI-specific considerations, emphasizing **trustworthiness**, **scenario-based testing**, **data quality**, **layered safety**

argumentation, and **continuous assurance**. The methodology employs the V-model combined with a W-model for learning assurance, defines a multimodal testing strategy (virtual, XiL, field), and frames certification around a layered safety case. It highlights the importance of data provenance, bias mitigation, ethical compliance, and human oversight, recognizing that certification must address not only technical performance but also societal and regulatory expectations. The report recommends horizontal standardization and modular ODD definitions to maximize synergies with other domains and facilitate reuse of components. 4.7 Safe AI using a driverless regional train as an example

3.7 THE COMMUNITY TO ACCELERATE THE DEPLOYMENT OF RESPONSIBLE AI IN INDUSTRIAL SYSTEMS

3.7.1 Introduction and overview of Confiance.ai

Confiance.ai^[15] started in 2021 as the technology pillar of the French **Grand Défi** on securing, certifying and enhancing the reliability of AI-based systems. It delivered a four-year research and development program that concluded in September 2024 and has since evolved into the **European Trustworthy AI Foundation**, pan European community aimed at scaling trustworthy AI practices across industrial sectors. The program's mission is to **federate industrial, academic and scientific players around a common objective: accelerating the adoption of responsible AI in safety-critical and business-critical systems** (confiance.ai).

The core idea behind Confiance.ai is that trustworthy AI cannot be achieved by isolated tools or standalone guidelines. Instead, it requires a **holistic engineering framework** that couples system engineering practices with machine learning specific activities; it requires a **knowledge platform** that brings together methods, standards and tools; and it requires a **community** that keeps the framework UpToDate as technology and regulations evolve. This chapter examines how Confiance.ai has organized its community, the methodological backbone known as the Body of Knowledge (BoK), the associated knowledge graph, and the relevance of these assets to the R2DATO WP6.4 objectives.

3.7.1.1 Community and governance

Following the program phase, the Confiance.ai community has taken shape as a **European Trustworthy AI Foundation**. The foundation offers an **open-source ecosystem** that federates manufacturers, researchers and other stakeholders, providing a collaborative framework to develop and share trusted solutions tailored to industrial needs. The community pursues research into emerging AI topics – such as generative AI, hybrid AI, human-machine collaboration and AI/cybersecurity – and supports industry initiatives that embed Confiance.ai results into market-ready tooling. Membership provides access to a growing catalogue of open source components, to training sessions, and to working groups that transform Confiance.ai outputs into standards and certification practices.

The community today includes **large groups, research centers, universities, startups and standards organizations**, illustrating its multisectoral reach confiance.ai. According to a 2024 report, around **fifty partners** were involved, spanning sectors from aeronautics to

energy and mobility. The community sees itself as a driver of the global trustworthy-AI ecosystem and actively collaborates with international peers in Canada, Germany, and Australia. Its ambitions for the post-2024 period are threefold: **expand** the methodological scope (e.g., hybrid and generative AI, cybersecurity of AI in critical systems); **sustain** the initiative through a long-term foundation that federates actors around a shared roadmap; and **industrialize** the programme's results to help companies integrate trustworthy AI at scale, including tailoring tools to specific data typologies and use cases (sitsi.pacanalyst.com).

3.7.2 The Body of Knowledge (BoK)

3.7.2.1 Purpose and scope

A keystone deliverable of Confiance.ai is the **Body of Knowledge**. According to the Confiance.ai technology portal, the BoK **centralises essential knowledge on the principles and practices of trustworthy AI**. It covers the **end-to-end development process** of a system whose functionality is automated through machine learning while guaranteeing a high level of trust. Specifically, the BoK addresses challenges such as data engineering, robustness of ML models, explainability, embeddability, uncertainty quantification and monitoring. It also states that defining the Operational Design Domain (ODD) is a multi-sectorial concept that also should be applied in other domains. The BoK also is establishing a global verification and validation plan based on assurance cases. The repository is designed not only to collect knowledge but also to **offer resources to support engineering** (methods, tools), **in order to serve as a reference for professionals, researchers and students**, and **promote the work of the Confiance.ai program**.

3.7.2.2 Methodological structure – the “V-W” lifecycle

The BoK formalizes an **engineering method** that combines classic systems engineering (the “V” model) with a software engineering “W” cycle tailored to machine learning. Traditional “V” phases cover operational analysis, system specification, architectural design, integration, verification and validation. In the Confiance.ai method, two extra engineering domains – **ML-algorithm engineering** and **data engineering** – create a second “W” cycle nested within the “V”. This leads to a **15-step end-to-end lifecycle** for critical ML-based systems (summarized below). While the dynamic BoK website lists these activities in detail, the high-level sequence presented here is reconstructed from the Confiance.ai documentation and research articles and aligns with the activity labels used in the BoK:

1. **Perform Operational Analysis** to determine the intended purpose and automation objectives.
2. **Define the Operational Design Domain (ODD)** for the automated feature.
3. **Perform System Specification** of the feature.
4. **Perform Architecture Definition & System Design**.
5. **Specify & Architect the ML component**.
6. **Specify & Architect the ML model and its complementary software items**.
7. **Develop or Acquire the ML model & complementary software items**.
8. **Evaluate the trained ML model & complementary items**.

9. **Implement & Integrate the ML model** (and complementary items) in the target hardware (embedded software).
10. **Evaluate the implemented ML software** on the target hardware.
11. **Evaluate the ML component** within its integration context.
12. **Integrate the ML component into the system.**
13. **Verify the ML-based automated feature** at system level.
14. **Validate the ML-based automated feature** against the intended purpose and automation objectives.
15. **Deploy and Maintain** the ML-based system in operation.

These activities are augmented by **crosscutting concerns** such as data governance, stakeholder alignment, traceability to the EU AI Act, assurance case development, risk analysis, robustness evaluation, explainability, monitoring, uncertainty quantification and embeddability. They ensure that evidence generated in earlier phases is reused during integration, verification, validation and postdeployment monitoring. The BoK offers **role-based views** for systems engineers, data engineers, ML engineers, integration/verification engineers and embedded software engineers so that each stakeholder can see only the tasks, inputs and expected deliverables relevant to them.

3.7.2.3 Functional sets and tools

The BoK is not merely a textual guideline; it also indexes a **tool stack** organized into **nine functional sets**. A 2024 analysis of Confiance.ai emphasized that the tool stack is evolutionary and includes both proprietary and opensource solutions (sitsi.pacanalyst.com). The nine functional sets correspond to specific engineering processes, ranging from **end-to-end engineering**, **data lifecycle management**, **model and component lifecycle management**, to **evaluation of robustness**, **explainability** and other trustworthiness attributes (sitsi.pacanalyst.com). This modular structure lets industrial users “drop in” appropriate tools at the right phase of the lifecycle, ensuring that compliance and performance requirements are addressed systematically.

3.7.2.4 Relevance for WP6.4 and future work based on WP6.4

For R2DATO’s WP6.4 and future work on perception certification of functions optionally using AI algorithms, the BoK offers several advantages:

- **Readymade structure for pilot documentation.** By adopting the BoK’s activity model, future work based on findings of WP6.4 can support flagship area 2 pilots across the entire VW cycle, ensuring consistency in how operational analysis, ODD definition, ML model development and integration are handled.
- **Application of the EU AI Act.** The BoK addresses high-risk AI requirements such as robustness, explainability, data quality and uncertainty, offering concrete processes and tools that can serve as evidence during conformity assessments (confiance.ai).
- **Integration with assurance cases.** Each BoK activity links to expected artefacts (e.g., requirements, models, test reports) that feed into structured assurance cases.

- **Modular tool integration.** The functional sets and catalogue provide a menu of software components (for robustness evaluation, explainability, synthetic data, etc.)..

3.7.3 Activities for engineering a critical ML-based system

Although the BoK website hosts an interactive representation of the lifecycle, its underlying structure can be described using the combination of systems engineering and ML-specific processes. In summary, the **life cycle shown in Figure 5 begins with understanding the operational context** (activities 1–4) by analyzing the intended purpose of the automation, defining the ODD and specifying the system architecture. It then transitions into the **ML development “W” cycle** (activities 5–10), where the ML component and model are specified, developed, evaluated and implemented on the target hardware. The subsequent activities (11–15) **focus on integration, verification, validation and maintenance** in real-world operation. This end-to-end view ensures that quality attributes such as safety, robustness and explainability are addressed **from concept to deployment**, rather than being bolted on at the end.

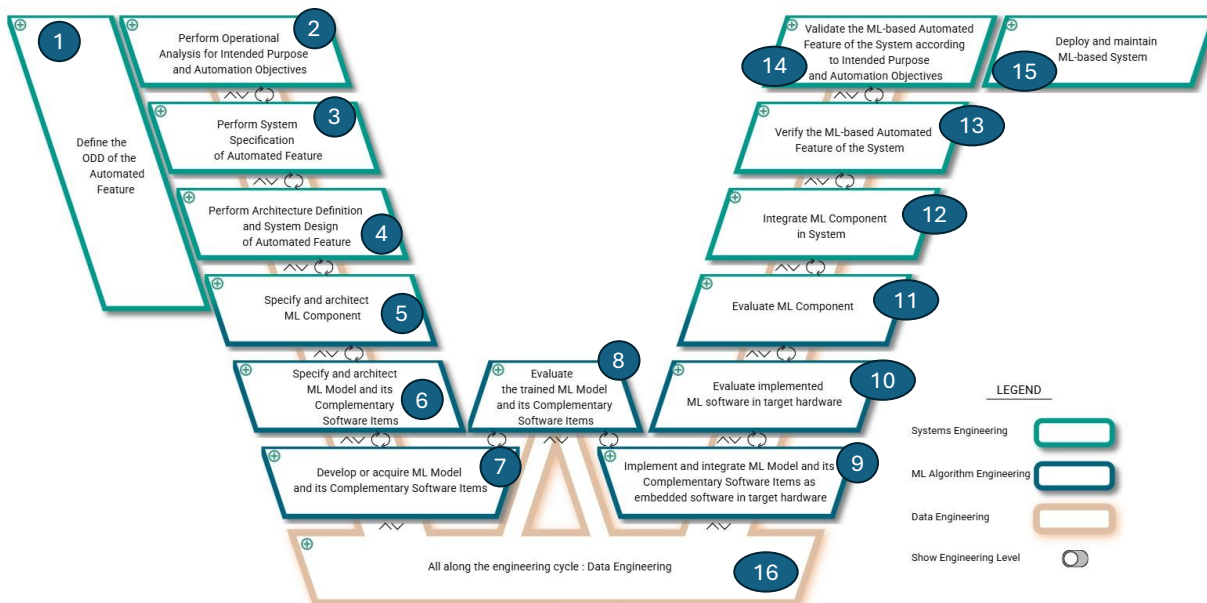


Figure 5: "Activities for the engineering of a critical ML-based system" according to Confiance.ai [Body Of Knowledge] [15] with number in ellipses to indicate the referenced number of the phase.

A notable aspect of Confiance.ai’s approach is the **explicit separation between model artefacts and software artefacts**. Activities 5 and 6 make a clear distinction between specifying the ML component (the logical interface and non-ML code that surrounds the model) and specifying the ML model itself (the trainable model and its hyperparameters). This separation is essential for certification bodies to trace how changes in data or model parameters influence system behavior. The later activities distinguish between evaluating the **trained model** (activity 8), evaluating its **embedded implementation** (activity 10) and

evaluating the **ML component** as part of the system (activity 11), ensuring that performance is checked at every level of abstraction.

The BoK also insists that **ML development must incorporate data governance and risk analysis from the outset**. Data engineering is treated as its own discipline, covering data collection, quality assessment, annotation, balancing, augmentation and synthetic data generation. Risk analysis – including hazard analysis, failure mode and effects analysis (FMEA), adversarial threat modelling and uncertainty quantification – runs in parallel to the technical development activities. Finally, **post deployment monitoring** is an integral activity: the BoK prescribes mechanisms to detect data distribution drift, monitor performance degradation and trigger model updates, aligning with upcoming regulatory obligations for high-risk AI systems.

3.7.4 The Confiance.ai knowledge graph

3.7.4.1 Purpose and implementation

To make the BoK and its taxonomy machine actionable, Confiance.ai has developed a **Knowledge Graph** (KG). According to the Knowledge Hub portal, the KG uses web semantic standard languages such as **RDF, RDFS and OWL**; this choice allows the integration of the graph with other semantic web tooling and enables export in **RDF/XML, Turtle, NTriples and JSONLD** formats (base.knowledgehub.confiance.ai). Unstructured data from research departments (PDF documents, CSV tables, BibTeX files) is syntactically analysed and converted into structured data to create **a single source of truth** for the ontology (base.knowledgehub.confiance.ai). The graph is stored in **Blazegraph**, an opensource triplestore that supports largescale graph queries (base.knowledgehub.confiance.ai).

3.7.4.2 Content layers

The knowledge graph is built around multiple layers, each capturing a different dimension of the BoK:

Layer	Representative entities	Possible queries (illustrative)
Taxonomy	trustworthiness attributes (robustness, explainability, uncertainty), life-cycle activities (Perform Operational Analysis, Define ODD, etc.), artefact types (requirements, models, test reports)	Example query: AUTOMATION_OBJECTIVE → affected_risk → ROBUSTNESS – asks which risks are associated with a given automation objective.
Methods	definitions of ODD workflows, assurance case patterns, robustness test procedures	Example query: HAS_STEP some ODD_INITIALISATION – retrieves processes that include an ODD initiation step.

Layer	Representative entities	Possible queries (illustrative)
Tools & Components	Robustness measurement libraries, explainability toolkits, data quality pipelines	Example query: COMPONENT provides_metric AccuracyImprovement – returns components that improve a given metric.
Standards mapping	ISO/IEC 5338 clauses, SAE ARP 6983 process rules, EU AI Act requirements	Example query: STANDARD_CLAUSE 4.3 fulfilled_by Activity#EvaluateMLComponent – maps a standard clause to the corresponding BoK activity.
Evidence graph	relationships between claims, arguments and artefacts used in assurance cases	Example query: ARGUMENT supports Claim#PerformanceUnderODD – finds arguments that support a claim about system performance under the defined ODD.

The KG can thus support **semantic search** across engineering activities, tools, standards and evidence..

3.7.4.3 Access and exploration

The Knowledge Hub provides a **Graph Explorer** that allows users to query and visualize the KG. Because the graph uses standard semantic web languages, it can be accessed through SPARQL endpoints and integrated with other graphs or knowledge management systems. The portal documentation recommends using the graph to explore relationships between trustworthiness attributes, engineering tasks and available tools. For instance, an engineer interested in **explainability** can query which BoK activities require explainability artefacts and which tools provide such functions. This capability moves the BoK beyond a static repository and positions it as a living, computationally exploitable knowledge base.

3.7.5 Implications for WP6.4 and recommendations

The Confiance.ai community and its BoK/knowledge graph infrastructure have direct relevance to R2DATO's WP6.4, which focuses on deploying responsible AI in industrial robotic systems. Below we outline major implications and recommended actions.

3.7.5.1 Alignment with regulations and standards

The BoK explicitly addresses challenges highlighted by emerging **AI regulations**. It covers **data quality**, **robustness**, **explainability**, **uncertainty quantification** and **continuous monitoring** confiance.ai, which correspond to the high-risk requirements in the EU AI Act. Future work based on WP6.4 can use BoK activities and artefact templates to demonstrate compliance and build **assurance cases** that show how system-level requirements and data-level evidence fulfil regulatory obligations. The BoK's mapping to standards (ISO/IEC 5338, SAE ARP 6983) ensures that the projects following the BoK lifecycle as proposed for WP6.4 based activities are aligned with the wider certification ecosystem.

3.7.5.2 Reusable engineering stack and tools

By adopting the BoK's **V-W lifecycle** and integrating its **functional sets**, WP6.4 pilots can avoid reinventing the wheel. The BoK's role-based views help assign responsibilities across system engineers, data scientists and software developers. The tool catalogue includes components for robustness measurement, explainability, synthetic data generation, embeddability and monitoring. WP6.4 should pilot these tools on at least one use case to evaluate their maturity, integration effort and suitability for robotics. Feedback from these pilots can be channelled back to the Confiance.ai community to refine the tools and methods.

3.7.5.3 Knowledge-centric tool-chain

Connecting WP6.4's data space to the Confiance.ai knowledge graph offers several advantages. Semantic integration would allow cross-querying between R2DATO artefacts (e.g., robotic system requirements, data sets, test reports) and the BoK ontology. Automated gap analysis could detect missing assurance artefacts or incomplete requirement flows. The graph could also be used to generate evidence packages for auditors, linking claims to supporting data and tests. Establishing a bridge ontology between R2DATO and Confiance.ai is therefore a recommended priority.

3.7.6 Conclusion

Confiance.ai has transitioned from a national research program into a **pan-European community** dedicated to responsible AI in industrial systems. Its **Body of Knowledge** provides a structured, trustworthiness-oriented method that couples classical systems engineering with ML-specific development to form a comprehensive "**VW**" **lifecycle**. The BoK is complemented by **functional tool sets** and a **knowledge graph** implemented using semantic-web standards base.knowledgehub.confiance.ai. Together, these resources address high-risk AI requirements such as data quality, robustness, explainability and continuous monitoring confiance.ai.

For R2DATO's WP6.4, adopting and extending the Confiance.ai approach offers clear benefits: it provides ready-made engineering structures, supports regulatory compliance, enables semantic integration through the knowledge graph.

3.8 SAFE AI USING A DRIVERLESS REGIONAL TRAIN AS AN EXAMPLE

3.8.1 Purpose of the safe.trAI_n demonstration

The **safe.trAI_n**[11][12][13][14] project investigated how perception systems based on artificial intelligence (AI) could be used to enable fully automated operation of regional trains (Grade-of-Automation 3/4). While the demonstrator train itself is a regional train, the project findings were deliberately aligned with R2DATO's automation ambition and provide an example of how a certification process for perception-based functions could be designed. The safe.trAI_n study established acceptance criteria for the safety function, developed the **Landscape of AI Safety Concerns (LAISC)**, formulated a structured certification argument

(five pillars of the safety-case strategy), implemented a modular perception architecture with built-in redundancy and monitoring, and proposed an end-to-end development and validation process named safeMLOps (see Figure 15). This chapter summarizes the safe.trAln results relevant to WP 6.4, adhering closely to the content of the project report. A detailed description is available in German ([Safe.trAln - Sichere KI am Beispiel fahrerloser Regionalzug](#)).

3.8.2 Acceptance criteria for the safety function

3.8.2.1 Motivation for quantitative safety targets

In early hazard analyses performed within safe.trAln[11], experts noted that regional passenger trains rarely encounter dangerous obstacles (less than once per year). Consequently, the project adopted the Probability of Failure on Demand (PFD) as the main acceptance criterion for the perception system. The PFD measures the probability that the system fails when a demand for obstacle detection occurs. Based on risk analyses using event trees, a quantitative safety target of 1 % PFD was derived, corresponding to Safety Integrity Level 2 (SIL 2), which aligns with the ATO Risk study.

3.8.2.2 Recall as the operative metric

The project formalized the PFD using the recall metric familiar from machine-learning evaluation:

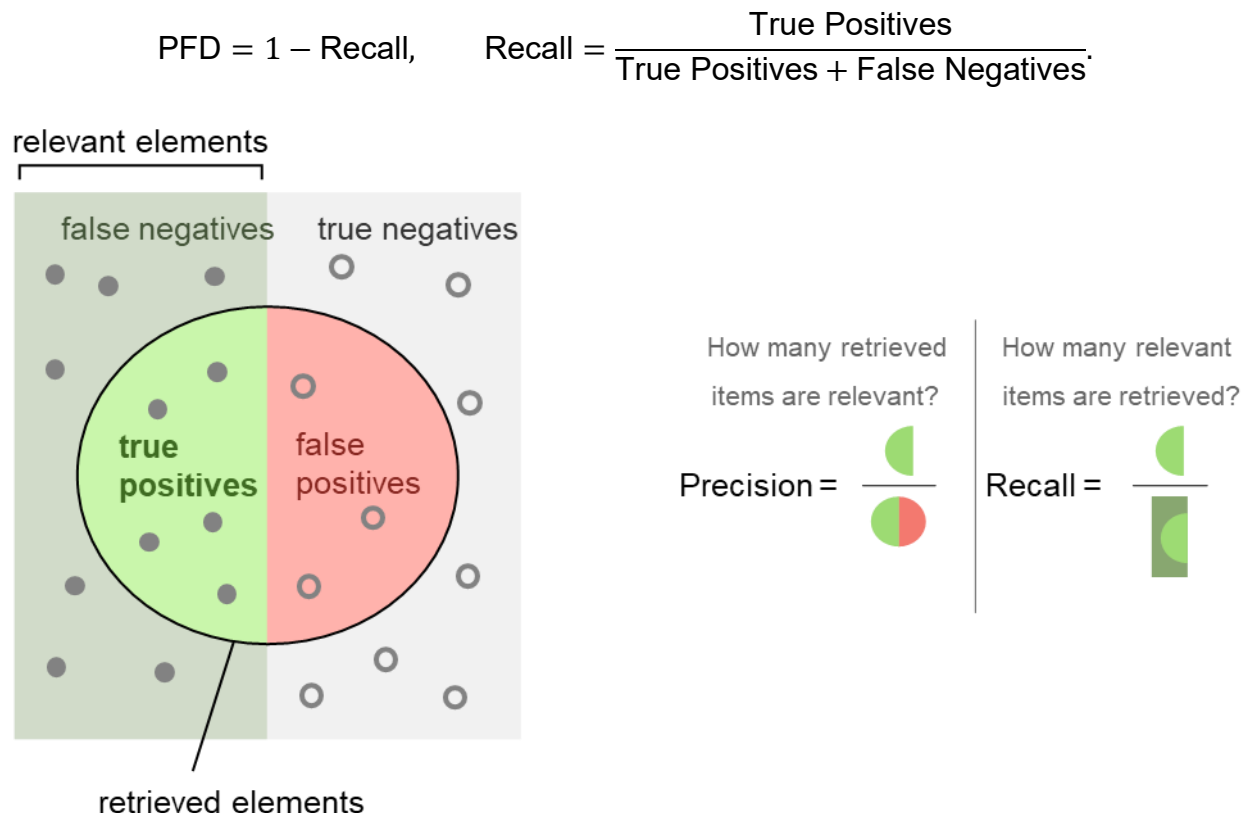


Figure 6: Definition of important metrics used in Statistics and machine learning as used in the safe.trAln project.

In the context of object detection, true positives correspond to obstacles correctly detected, whereas false negatives represent missed detections. Figure 21 in the safe.trAln report illustrates the relationship between recall, precision and regions representing true positives, true negatives, false positives and false negatives. By targeting a recall such that the PFD does not exceed 1 %, the project derived acceptance criteria that directly link the quantitative safety goal to machine-learning performance. This metric allows testers to assess whether a perception algorithm meets the required SIL target and provides a clear acceptance criterion for certification.

3.8.2.3 Risk zonation and reaction requirements

The acceptance criteria are not applied uniformly across all objects. The project identified different risk zones in front of the train and associated detection requirements (“Area of Interest”, “warning area” and “danger area”) together with a black list (objects requiring reaction) and white list (objects that do not require reaction). Such zonation informs the definition of the PFD and recall metrics by prioritising critical scenarios.

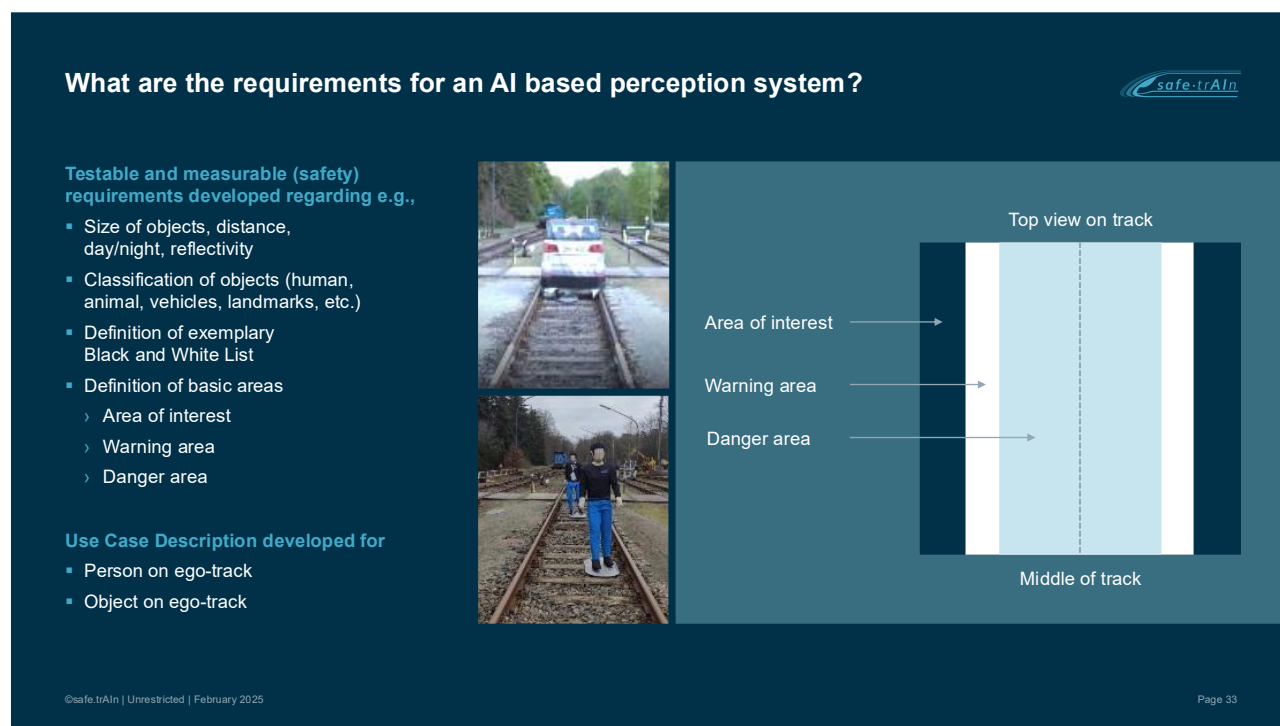


Figure 7: Definition of the requirement for an AI based perception system according to the safe.trAln project.

3.8.3 Landscape of AI Safety Concerns (LAISC)

3.8.3.1 Concept and purpose

To bridge the gap between AI-based systems and conventional safety standards, safe.trAln developed the Landscape of AI Safety Concerns (LAISC). The LAISC is a structured catalogue of AI-specific hazards and potential mitigation measures covering all phases of the AI life-cycle. The report emphasises that “an AI Safety Concern is an AI-specific issue that can negatively affect the safety of a system”. Working with project partners

(Siemens AG, Fraunhofer IKS and Siemens Mobility), the consortium identified 22 AI Safety Concerns through extensive state-of-the-art analyses. Figure 22 (reproduced below) visualises these concerns and groups them into categories such as inadequate specification of the operational design domain (ODD), insufficient AI development documentation, AI-related hardware issues, choice of untrustworthy data sources, inadequate data representation and concept drift.

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

Figure 8: Landscape of AI Safety Concerns according to the safe.trAIn project. The safe.trAIn report identifies 22 AI-specific safety concerns and organises them in categories such as specification of the ODD, data quality, robustness, transparency, and integration issues.

For each AI Safety Concern, the project derived **verifiable requirements**—criteria that demonstrate either that a specific hazard cannot occur or that it is effectively controlled. Metrics and evidences are then used to provide proof that these requirements have been met. This structure allows certification bodies to trace hazards to concrete verification artefacts.

3.8.3.2 Four-step LAISC application process

The safe.trAIn consortium formulated a four-step process for applying the LAISC:

1. **Identification of AI Safety Concerns:** Collect all relevant AI safety concerns based on state-of-the-art research and project experience.
2. **Use-case-specific decomposition:** Tailor the generic concerns to the specific use case, providing actionable details. This step ensures that the derived safety management strategies are relevant and effective.
3. **Derivation of verifiable requirements:** Define criteria for each AI Safety Concern that allow certification authorities to determine when the concern has been satisfactorily addressed.
4. **Application of metrics and mitigation measures throughout the AI life-cycle:** Generate evidence across development, testing and operation to demonstrate that mitigations are effective and that hazards remain controlled. This continuous validation ensures transparency and traceability across the life-cycle.

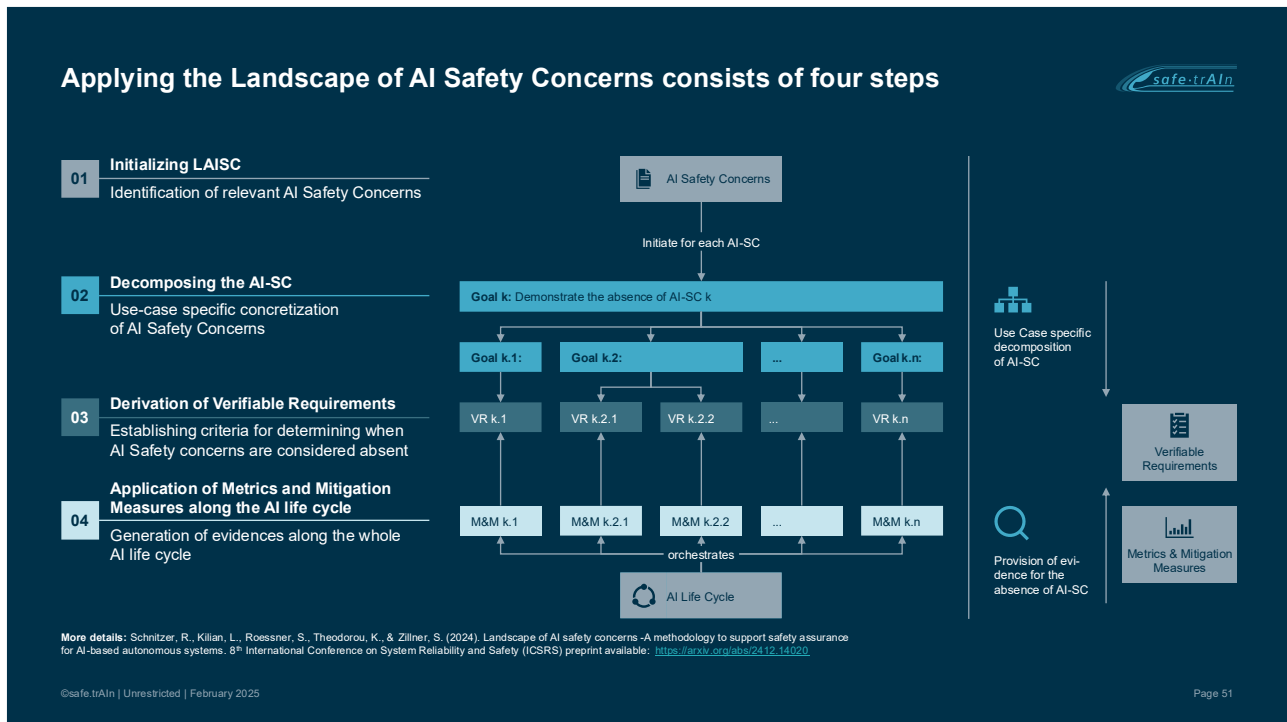


Figure 9 Depicts this four-step process graphically and emphasises that the process ensures transparency and traceability. An example of the LAISC application is provided for the “Inaccurate Data Labels” concern, where goals such as ensuring accurate data labels, empirically validating their impact on performance and automatically detecting and replacing inaccurate labels are broken down into verifiable requirements.

3.8.3.3 Integration of LAISC into the certification process

The LAISC is not an isolated list but a foundational element of the certification argument. The safety case developed in safe.trAIIn uses the LAISC to ensure that AI-specific hazards are systematically addressed. In particular, the LAISC informs the development process (first pillar of the safety case) and the component-level safety argumentation using Goal Structuring Notation (GSN) (see Section 4.7.5.3). By providing verifiable requirements, metrics and mitigation measures, the LAISC bridges the gap between high-level hazards and concrete evidences needed for certification.

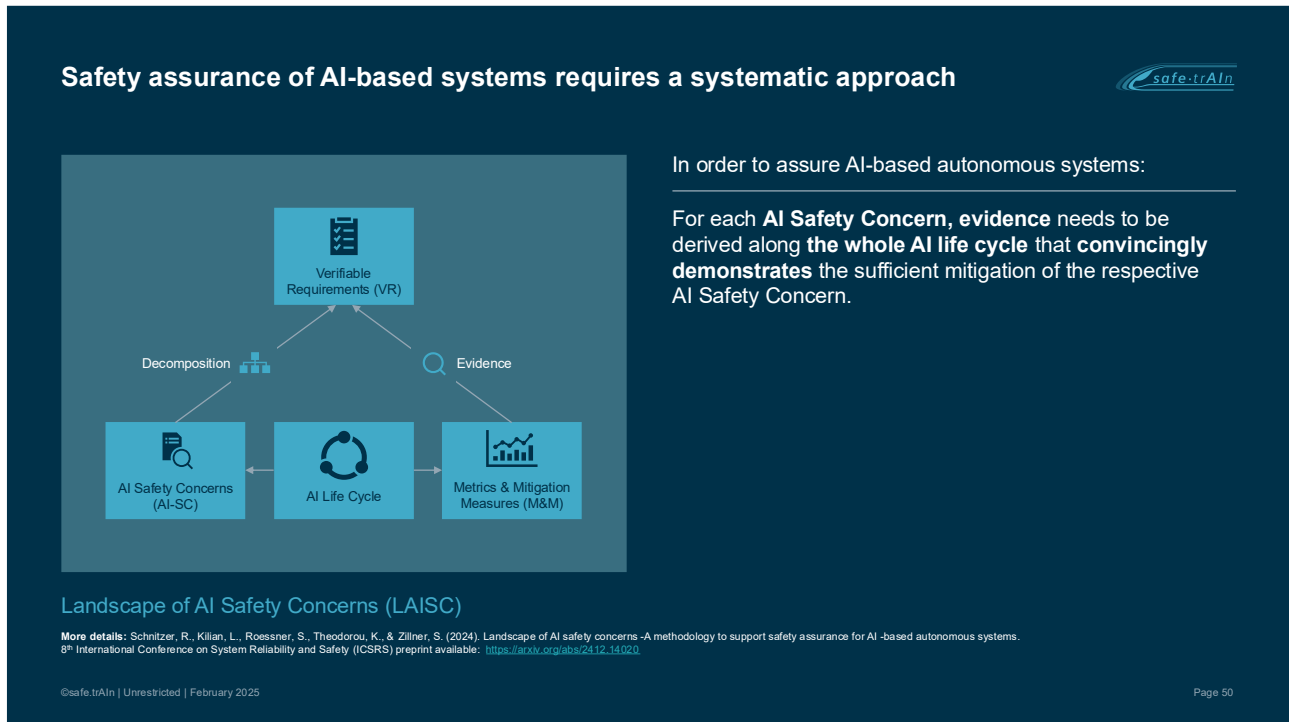


Figure 10: Safety assurance of AI-based systems requires a systematic approach according to the safe.trAln project.

3.8.4 Safe perception system architecture

3.8.4.1 Principles of safe object detection

The perception system in safe.trAln combines multiple sensor modalities (camera, LiDAR) and employs both AI-based and conventional object detectors. Figure 11 illustrates the architecture for **safe object detection**. It shows dissimilar detection paths: AI-based detectors for camera and LiDAR streams operate in parallel with conventional algorithms; their outputs are monitored before fusion, and a separate module fuses the results to determine obstacles. Key principles include:

- **Dissimilarity:** Use of heterogeneous algorithms and sensor modalities reduces common-cause failures.
- **Redundant detection:** Parallel detectors provide alternative estimates, and monitor components validate the plausibility of each result.
- **Uncertainty estimation:** Each detector produces uncertainty values that are propagated through the fusion process.

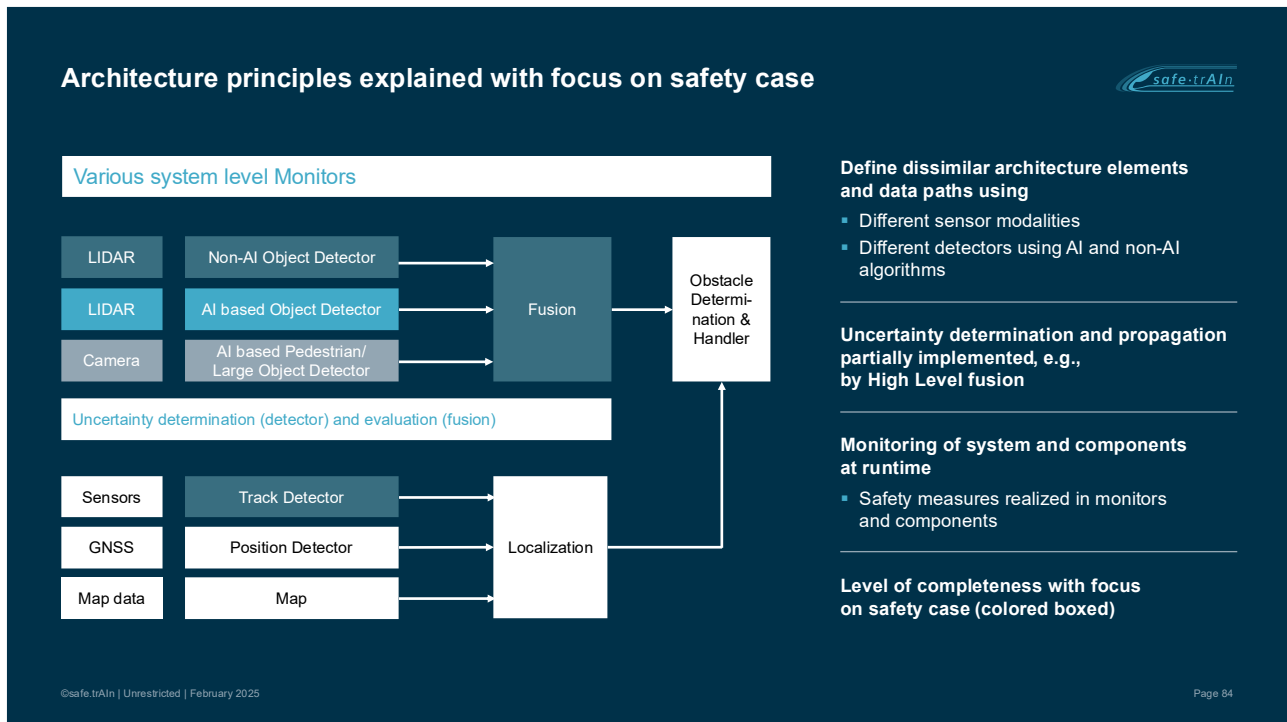


Figure 11: Principles of a safe object-detection architecture. AI-based and conventional detectors process camera and LiDAR data in parallel. Monitors validate each detector's output, and a fusion module combines the results to determine obstacles.

3.8.4.2 Safe sensor fusion

Object detection feeds into a **sensor-fusion architecture** that is structured into three levels: low-level fusion combines raw sensor measurements, mid-level fusion combines object-level detections, and high-level fusion constructs an environment model. Figure 12 illustrates this multi-layer architecture. The fusion layers are supervised by a **supervisor** that monitors uncertainties and ensures consistency, while a dedicated uncertainty determination module quantifies and propagates uncertainties across the pipeline. Key elements are:

- **Heterogeneous sensors and algorithms:** Fusion combines information from different sensor types and detection methods.
- **Uncertainty propagation:** Uncertainties are determined at each level and used by the supervisor to detect anomalies or inconsistent data.
- **Monitoring:** Dedicated monitors watch the outputs of fusion and detection components and trigger safe reactions if inconsistencies arise.

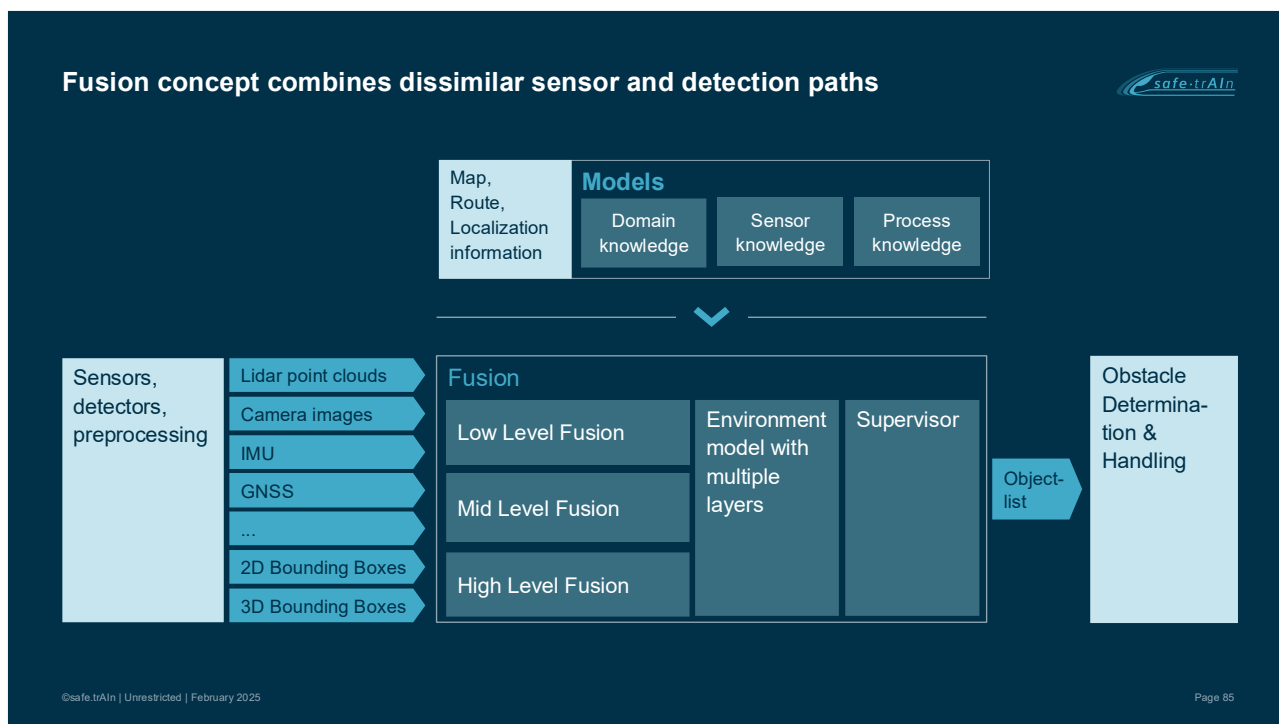


Figure 12: Concepts in the architecture for safe sensor fusion. Sensor data are combined across low-, mid- and high-level fusion layers; a supervisor monitors uncertainties and ensures consistency.

3.8.4.3 Training data and risk zonation

Training the AI components relied on open datasets such as **OSDaR23** (Open Sensor Data for Rail 2023) and proprietary data collected by Siemens Mobility. These datasets were used to train foundation models for large obstacle detection and other perception tasks. The architecture also implements risk zones (area of interest, warning area, danger area) and object categories (black-list and white-list) to prioritise detection and reaction requirements.

3.8.5 Safety-case strategy for the GoA3/4 architecture

3.8.5.1 Five pillars of the safety-case strategy

Based on the safe perception architecture, the project developed a safety-case strategy composed of **five pillars**, depicted in Figure 13.

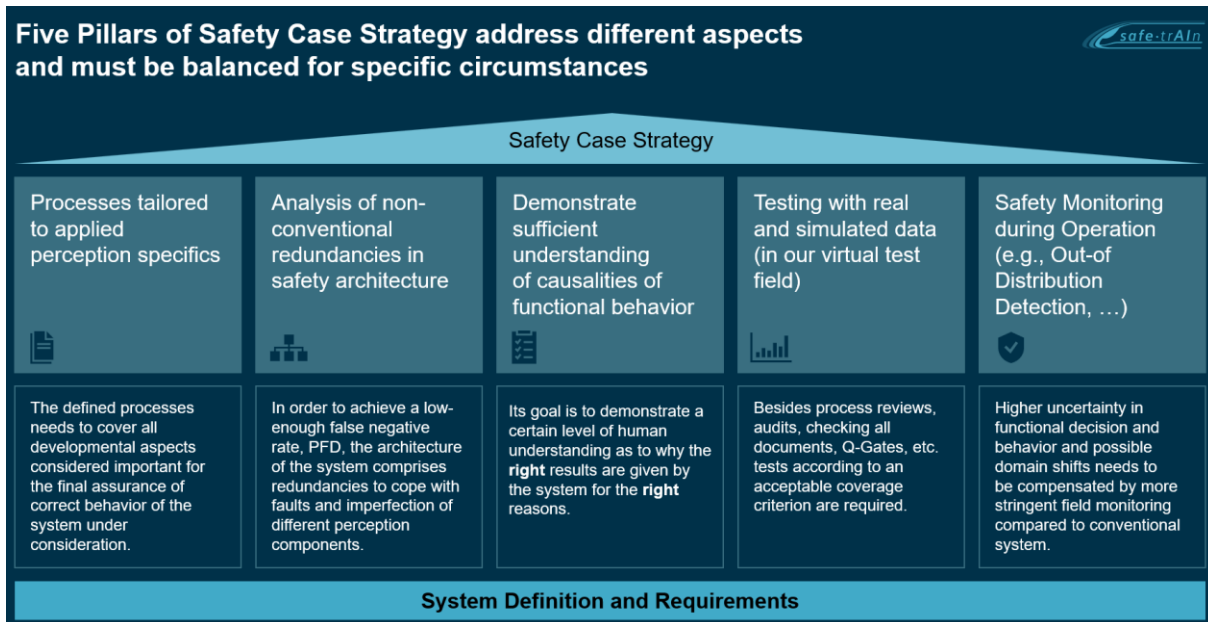


Figure 13: Five pillars of the safety-case strategy. The pillars address development processes, non-conventional redundancies, understanding of causalities, testing, and safety monitoring.

1. **Processes tailored to applied perception specifics:** The development process must be adapted to the perception task and cover all aspects important for the safety argumentation. The **Operational Design Domain (ODD)** and the LAISC play central roles. Each AI Safety Concern is addressed with convincing evidences over the entire AI life-cycle.
2. **Analysis of non-conventional redundancies:** The safety architecture integrates **innovative redundancies** beyond identical hardware duplication. Different sensor modalities, detection methods (AI-based and conventional) and data-processing paths are combined to provide comprehensive coverage and fault tolerance. A critical aspect of this strategy is the **determination and evaluation of uncertainties** to ensure that trust levels of each sensor and detection path are continuously assessed. The concept of **monitors** protects both AI-based and conventional components and ensures they work together effectively.
3. **Demonstrate sufficient understanding of causalities of functional behaviour:** It must be shown that the system produces the right results for the right reasons. Close collaboration between AI and domain experts is essential. Transparency and trust in the system's decision-making must be established by analysing why the system makes certain decisions. While full end-to-end explainability is unrealistic, components must offer sufficient observability and explainability, for example through **saliency maps**. Understanding causalities improves stakeholder trust and helps identify potential biases or confounders.
4. **Testing with real and simulated data:** A comprehensive testing concept with adequate coverage is required to generate safety evidences. Each test level focuses on a specific test object and is supported by an appropriate test environment, including the virtual testbed developed in safe.trAI. Tests were performed at component and system level, with both real and synthetic data.

5. **Safety monitoring during operation:** Continuous monitoring ensures that the system remains safe during operation. The safe.trAIIn safety-case emphasises runtime monitoring methods such as out-of-distribution (OoD) detection to differentiate between valid objects and background, addressing variations in object occurrence probabilities. Monitoring is integral not only for initial certification but also for maintaining safety over time.

The safety-case strategy highlights a detailed understanding of functional behaviour and causal mechanisms, supporting transparency and justification of AI decisions. It underscores the need for **Failure Mode and Effects Analyses (FMEA)** at component level early in development, helping identify potential causes of failure and define necessary mitigation measures. In safe.trAIIn, FMEAs were carried out for the camera sensor, mid-level fusion, high-level fusion and track-detection components. System-level FMEA complemented the component analyses and allowed the architecture to be iteratively refined with safety measures.

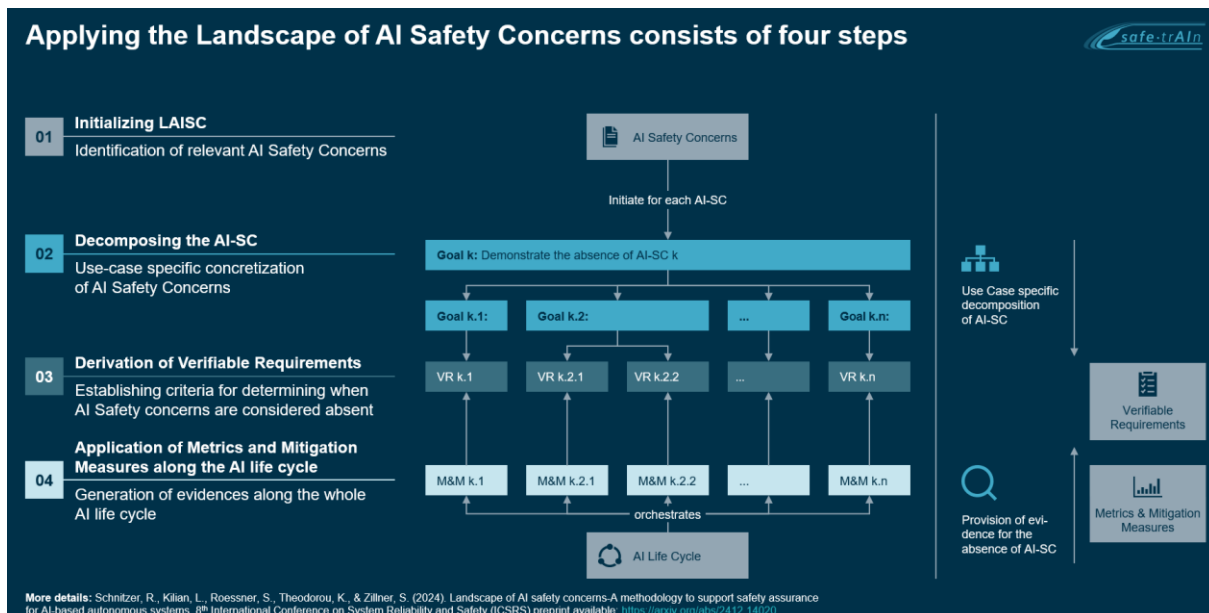


Figure 14: Process for the application of the "Landscape of AI safety concerns"

3.8.5.2 Formal safety-case concepts

The project adapted the **EN 50129** structure for system-level safety cases to incorporate AI/ML aspects. The safety-case document includes ML-specific aspects, results from component safety cases, and an annex table describing required steps for each AI life-cycle phase. Due to inherent uncertainties in AI, evidences are relative rather than absolute. The safety argument draws on supporting documents such as the architecture description, AI validation report, LAISC, safety validation concept and metric fact sheets.

At component level, Goal Structuring Notation (GSN) was used to graphically represent the safety argument. The tool **nLoop** (provided by Edge Case Research) was integrated with Siemens' data management platform to automatically generate validation reports and

3.8.5.3 Virtual test field and SafeMLOps

The diagram illustrates the Safety engineering lifecycle, which integrates System engineering, ML Life-Cycle, and Safety engineering across various stages of a product's lifecycle.

System engineering lifecycle (inspired by the development process of EN 50126-1) includes:

- System Design** (containing Domain Analysis, System Specification, and Funktional & Technical Architecture Description)
- System V&V** (containing Field Test, HW/SW Integration, and Virtual Field Test)
- SW (non-ML) Design & Implementation**
- ML/SW Integration**

Data & ML Life-cycle (inspired by the development process of EN 50126-1) includes:

- Data Acquisition & Preparation**
- ML Model Design**
- ML Model Training & Verification**

Safety engineering lifecycle includes:

- Hazard & Risk Analysis**
- Failure & Deficiency Analysis**
- System Operation in the field**
- Decommissioning/Retirement**
- Safety Assessment**

Key Milestones and Phases:

- Safety Concept Design** (encompassing Hazard & Risk Analysis, Failure & Deficiency Analysis, and System Design)
- Safety Case Development** (encompassing System V&V, Field Test, HW/SW Integration, Virtual Field Test, and Safety Assessment)

Legend:

- System engineering (Blue)
- ML Life-Cycle (Grey)
- Safety engineering (White)
- Out of scope in safe trAln (Dashed border)

The project also developed a **SafeMLOps** process that adapts continuous integration/continuous deployment principles to AI applications. Every change to the AI model or data triggers the entire assurance pipeline, from ODD and requirements definition through training and virtual testing to deployment and monitoring. The pipeline automatically updates the safety case when the model is adjusted, ensuring that safety analyses and test results remain consistent. This continuous development and safety process underpins a dynamic certification approach suitable for AI systems.

The safe.trAI demonstrator shows that a **driverless regional train** can be used as a concrete example to develop a certification process for perception systems. Key lessons for R2DATO WP 6.4 include:

- 29/09/2026

brake of the train may arise if a person is unable to leave the track in time or a large, heavy object is present in the critical area of the ego track of the train. This conclusion is conflicting with the results from ERJU R2DATO WP8.2 (See chapter 4.4) considering PER is operating in continuous mode according to safety standard, because PER has the task of detecting, reporting and continuously monitoring anomalies.

- **LAISC:** A comprehensive catalogue of AI safety concerns with verifiable requirements and metrics ensures that AI-specific hazards are systematically addressed.
- **Structured safety-case argument:** The five-pillar strategy covers development processes, redundancy analysis, causal understanding, testing and operational monitoring. It requires continuous evidences across the AI life-cycle and integrates component-level FMEA and GSN representations.
- **Continuous assurance:** The perception system, AI architecture, and supporting tool chain must be designed for continuous evolution. The system must be able to adapt to changes in the environment, such as the use of masks during the coronavirus pandemic or the market launch of e-scooters. Such changes mean continuous pressure for updates to object recognition systems. The SafeMLOps process demonstrates how certification may evolve into a continuous assurance approach where model and data updates automatically trigger reassessment and update of the safety case.

By adopting these principles, WP 6.4 can enhance the certification process for on-demand perception systems and ensure that AI-enabled automation functions meet stringent safety and reliability requirements.

4 INPUT IN THE PROCESS DESCRIPTION

4.1 REGULATORY AND NORMATIVE FRAMEWORK

4.1.1 Technical Specification for Interoperability

The regulatory TSI has been consulted as a support in the process description framework of the Perception System:

Reference	Document title	Last version
LOC&PAS TSI	Commission Regulation (EU) No 1302/2014 of 18 November 2014 amended by Regulation (EU) 2023/1694 of 10 August 2023 and (EU) 2025/675 of 4 April 2025	2025
Control Command and Signalling TSI	Commission Implementing Regulation (EU) 2023/1695 of 10 August 2023	2023
Safety in Railway Tunnels TSI	Commission Regulation (EU) No 1303/2014 of 18 November 2014 corrected by Regulation (EU) 2016/912 of 9 June 2016 and amended by Regulation (EU) 2019/776 of 16 May 2019	2019
Operation and Traffic Management TSI	Commission Implementing Regulation (EU) 2019/773 of 19 May 2019 amended by Implementing Regulation (EU) 2021/2238 of 15 December 2021 and Implementing Regulation (EU) 2023/1693	2023
Infrastructure TSI	Commission Regulation (EU) No 1299/2014 of 18 November 2014 amended by Implementing Regulation (EU) 2019/776 of 16 May 2019 and Implementing Regulation (EU) 2023/1694 of 10 August 2023	2023

The PER system must undergo conformity assessment procedures, involving a Notified Body, to verify compliance with the applicable TSIs. If requested, a Designated Body shall verify the conformity with respect to established national rules.

4.1.2 Common safety Methods Risk assessment

The safety framework is the set of application regulations used as mandatory for the safety analysis process. A safety case is typically worked out by the perception system supplier. An independent AssessmentBody shall undertake investigation to provide a judgement, based on evidence, of the suitability of the system under assessment to fulfil its safety requirements.

REF / Document ID	Document title	Version
CSM-RA 402/2013	COMMISSION IMPLEMENTING REGULATION (EU) No 402/2013 of 30 April 2013 on the common safety method for risk evaluation and assessment and repealing Regulation (EC) No 352/2009	402/2013
CSM-RA 2015/1136	COMMISSION IMPLEMENTING REGULATION (EU) 2015/1136 of 13 July 2015 amending Implementing Regulation (EU) No 402/2013 on the common safety method for risk evaluation and assessment	2015/1136

4.1.3 EU AI ACT

4.1.3.1 Scope and relevance for D6.4

The EU Artificial Intelligence Act (Regulation (EU) 2024/1689[20], “AI Act”) is a horizontal framework that applies across sectors and therefore also to AI-enabled perception (PER) functions in rolling stock if AI based components are used to implement PER. Within D6.4 it complements TSI-based conformity assessment and CSMRA safety assurance by introducing additional, AI-specific obligations (risk management, data and model governance, documentation, logging, transparency and human oversight, post-market monitoring). The section below refines the risk-based approach, explains how

the AI Act follows the New Legislative Framework (NLF), provides the staged applicability timeline, and preserves the note on the Commission’s Standardization Request—updated with the current (Oct 2025) status. The structure and tone align to the D6.4 layout in Chapter 4 (“Regulatory and normative Framework”).

4.1.3.2 Risk-based approach of the AI Act

The AI Act classifies AI uses by risk and couples each class with proportionate duties[20] (see also Figure 16):

- **Prohibited practices (Art. 5):** Certain uses are banned outright (e.g., social scoring of individuals in specified forms, some biometric categorization and deceptive manipulation techniques). These prohibitions have **early applicability** in the overall timeline (see 4.1.3.4). [EUR-Lex](#)
- **High-risk AI (HRAI):**
 - **Annex I linkage:** AI functions that are **safety components** of products or systems regulated by NLF-based Union harmonization legislation (e.g., product regulations listed in Annex I) fall in scope when those products are themselves subject to third-party conformity assessment.
 - **Annex III areas:** Stand-alone applications in sensitive domains (e.g., critical infrastructure) are high-risk when the use-case matches Annex III. In a rail context, PER functions integrated as safety components or used in critical infrastructure control can be high-risk and must fulfil Chapter III/Section 2 requirements (risk management, data governance, technical documentation, recording, transparency/human oversight, robustness, cybersecurity) plus conformity assessment under Art. 43. [EUR-Lex](#)
- **Limited-risk:** Transparency duties apply to specific features (e.g., AI systems that interact with humans or generate content) outside the prohibited/high-risk scopes. [EUR-Lex](#)
- **Minimal-risk:** No specific legal obligations—voluntary standards and good practice remain advisable. [EUR-Lex](#)

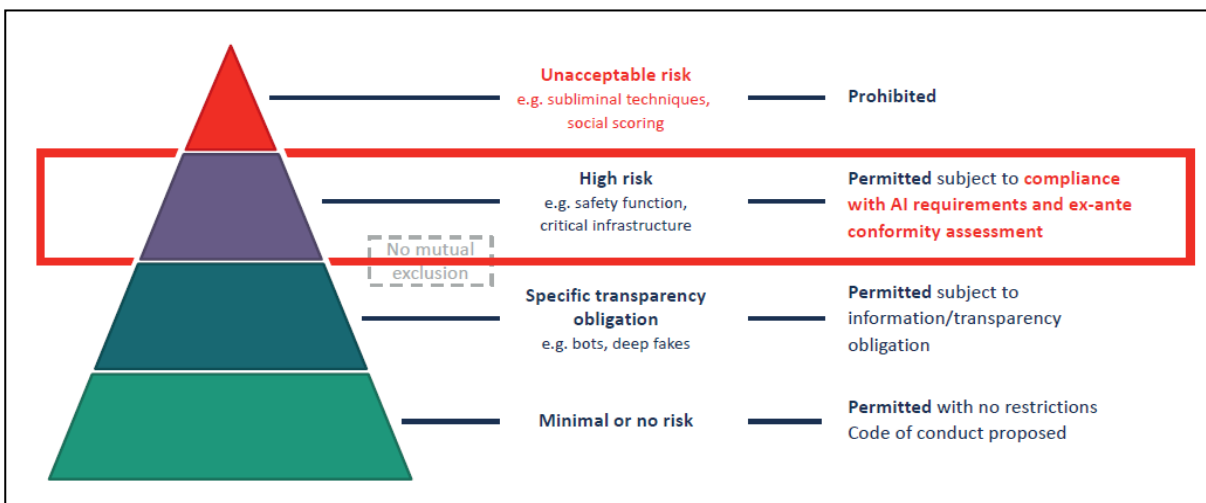


Figure 16: Risk-based approach of the AI Act: High-Risk class is potentially relevant for PER using AI

General-purpose AI (GPAI): The AI Act also regulates **GPAI models**, including additional obligations for **models with systemic risk** (e.g., reinforced testing, incident reporting, and information-sharing). The GPAI compliance phases are part of the staged timeline below (Aug 2025).

[Digital Strategy European Commission](#)

Implication for rail PER: If a PER function uses AI and is a safety component of a regulated product/subsystem or falls under Annex III, it is high-risk. The resulting obligations must be built into the D6.4 process (Stages 0–9), with traceable evidence from ODD definition through V&V to post-market monitoring.

4.1.3.3 The NLF architecture inside the AI Act

The AI Act follows the **New Legislative Framework (NLF)** model familiar from EU product law:

- **Essential requirements in law; technical solutions in standards:** Harmonized European standards (hENs) give **presumption of conformity** with the corresponding AI Act obligations once cited in the Official Journal (OJ). Where hENs are not yet available or adequate, the Commission may adopt **Common Specifications (CS)** (Art. 41). [EUR-Lex](#)
- **Conformity assessment (Art. 43; Annex VI/VII):** For high-risk AI, the law provides internal control (Annex VI) or quality-management/technical-documentation assessment involving a notified body (Annex VII), depending on the case and whether the relevant hENs are applied. [EUR-Lex](#)
- **Integration with sectoral law:** The Act **amends and interfaces** with sectoral regimes listed in Annex I (Section A) and explicitly **amends Directive (EU) 2016/797** on railway interoperability to ensure alignment where AI safety components are embedded in rail subsystems. This enables use of established sectoral pathways (TSIs, NoBo/DeBo) while adding AI-specific conditions. [EUR-Lex](#)

For background on the NLF (roles of economic operators, CE marking logic, modules of conformity assessment, and presumption of conformity), see the **Blue Guide (2022)**. [Internal Market & SMEs+1](#)

4.1.3.4 Assessment routes for rail

Two typical pathways are relevant:

- **Embedded HRAI in regulated products (Annex I reference):** Where PER with AI may be a safety component within a rail product/subsystem already subject to NLF legislation and TSI assessment, **AI-Act obligations are integrated** into the existing sectoral conformity and NoBo processes (Art. 43(3)), with AI-specific essential requirements demonstrated via hENs or CS and documented in the technical file. [EUR-Lex](#)
- **Stand-alone HRAI (Annex III use-cases):** Where PER with AI is deployed as a stand-alone high-risk application (e.g., critical-infrastructure operations or safety-relevant decision support), apply **Art. 43** directly using Annex VI or VII, depending on the standardization status and risk profile. [EUR-Lex](#)

In both cases, **CSMRA**[6] obligations for significant changes and **TSI-based** conformity continue to apply in parallel. D6.4 Stage-mapping below shows how to embed AI-Act duties into the existing rail process.

4.1.3.5 Phased applicability and key dates

- **OJ publication:** 12 July 2024; **entry into force:** **1 August 2024**. [EUR-Lex](#)
- **From 2 February 2025: Prohibitions** (Art. 5) and **AI-literacy** provisions apply. [Digital Strategy European Commission](#)
- **From 2 August 2025: GPAI model** obligations and governance rules apply. [Digital Strategy European Commission](#)
- **From 2 August 2026:** General **full applicability** for high-risk AI obligations, unless special transitions apply. [Digital Strategy European Commission](#)
- **Until 2 August 2027:** Extended transition for **high-risk AI embedded into regulated products**, to align AI-Act compliance with sectoral product conformity cycles. [Digital Strategy European Commission](#)

These fixed dates frame planning for pilots and market introduction of PER-equipped rolling stock.

4.1.3.6 Standardisation Request and current status (Oct 2025)

Mandate: The Commission issued **Implementing Decision C(2023) 3215** (May 2023) to CEN/CENELEC (with ETSI consultation) to draft harmonized standards in support of the AI Act (risk management, data-/model governance, QMS for AI, conformity assessment, robustness/cybersecurity, human oversight, post-market monitoring, logging, transparency). [Economics and Innovation Ministry](#)

Amendment & extended deadline: A **draft amending request** was published in **January 2025** to align the mandate with the final Act and to extend deadlines. Subsequent Commission steps resulted in a **new delivery date of 31 August 2025** for the requested packages. [European Commission+1](#)

Status (October 2025): According to a Council working-party note (12 September 2025), **CEN and CENELEC (JTC 21) did not meet** the 31 August 2025 deadline and indicated that work will **extend significantly beyond** that date; an updated timeline has been presented to Member States. This implies **delayed availability of OJ-cited harmonized standards** conferring presumption of conformity. [Data Consilium](#)

Consequence for projects: Until the relevant hENs are available and cited, manufacturers/operators must (i) rely on best-available standards and evidence without presumption, or (ii) apply **Common Specifications** if the Commission adopts them under **Art. 41** to bridge standardization gaps. Conformity routes may therefore prefer **Annex VII** (NoBo/QMS path) where appropriate, with careful documentation of “state of the art” justifications. [EUR-Lex](#)

Note – preservation of existing D6.4 content: This updated status **retains** the prior D6.4 note on the Standardisation Request (C(2023) 3215) and **replaces** the earlier date assumptions with the current information above. Ongoing monitoring of **CEN/CENELEC JTC 21** publications remains essential.

4.1.3.7 Practical mapping to the D6.4 workflow (Stages 0–9)

To keep the chapter consistent with the rest of D6.4, the table below maps **AI-Act obligations** to the **D6.4 process** (Sections 6.4.1–6.4.10). It is intended as a checklist for dossiers and planning.

- **Stage 0 – Initiation & governance:** Establish an **AI-Act compliance plan**; nominate roles; define **lifecycle risk management (Art. 9)** as a continuous process; plan **technical documentation (Art. 11)**, **record-keeping (Art. 12)**, and **post-market monitoring (Art. 61 ff.)** from day one. [EUR-Lex](#)
- **Stage 1–3 – ODD, HD-map, simulation foundations:** Implement **data/dataset governance (Art. 10)** tied to the **ODD**, with provenance, representativeness and bias controls; specify format/quality for HD-maps and simulation assets as safety evidence; pre-allocate document structure for the future **technical file**.
- **Stage 4 – Safety & hazard analysis (CSMRA):**[6] Integrate **AI-specific hazards** (e.g., data/label errors, drift, adversarial robustness) into the hazard log; specify **transparency & human oversight (Arts. 13–14)**, and safety metrics (accuracy, robustness, cybersecurity—Art. 15). [EUR-Lex](#)
- **Stage 5–6 – Verification & field validation:** Produce evidence for **robustness & performance (Art. 15)**, **logging/traceability (Art. 12)**, latency and availability within the ODD; correlate simulation with field data; confirm **human-oversight** and fallback behaviors. [EUR-Lex](#)
- **Stage 7 – Conformity & safety assessment:** Select **Art. 43** path:
 - Embedded HRAI in a regulated product → integrate AI-Act requirements into the **sectoral NoBo** assessment.
 - Stand-alone HRAI → **Annex VI** (internal control) or **Annex VII** (QMS/NoBo). Consider lack of hENs and potential **Common Specifications (Art. 41)** when planning the route. [EUR-Lex](#)
- **Stage 8–9 – Authorization, operation, change:** Activate **post-market monitoring & incident reporting**, including triggers for **significant change** and re-assessment under CSM-RA and Art. 43(4); link ODD updates and HD-map changes to controlled re-validation. [EUR-Lex](#)

4.1.3.8 What to watch (2025–2026)

- **Harmonized standards (CEN/CENELEC JTC 21):** Track delivery schedules and OJ citations that will enable presumption of conformity for HRAI obligations. Until then, plan with Annex VII and **state-of-the-art** justifications; monitor possible **Common Specifications** under Art. 41. [Data Consilium+1](#)
- **Rail interface (Directive 2016/797):** Follow ERA/Commission guidance on integrating AI-Act duties into TSI/NoBo processes as embedded HRAI become commonplace in rail subsystems. [EUR-Lex](#)

4.1.3.9 Summary for WP 6.4

- The AI Act **adds AI-specific obligations** on top of TSIs/CSM-RA for **high-risk** PER functions (either embedded or stand-alone). [EUR-Lex](#)
- It is **NLF-consistent**: demonstrate essential requirements via **harmonised standards** (when available) or **Common Specifications**; select **Art. 43** assessment path accordingly. [EUR-Lex](#)
- The **timeline is staged** (Feb 2025 prohibitions; Aug 2025 GPAI; Aug 2026 high-risk; special transition for embedded HRAI to Aug 2027). Plan project gates and evidence accordingly. [Digital Strategy European Commission](#)
- The **Standardisation Request (C(2023) 3215)** was amended and **extended to 31 Aug 2025**—but **JTC 21 did not meet** that deadline; work continues beyond it. Until OJ-cited hENs exist, rely on Annex VII and watch for **Art. 41 CS**. [European Commission+1](#)

4.1.4 EN 50716

EN 50716 Railway Applications – Requirements for software development (50128 & 50657 merge standard) has provided clarifications in §C.3 Artificial Intelligence and Machine Learning:

“In the context of applications in the scope of EN 50716, there are four challenges of machine learning which are hard in the sense that it is not yet possible to be confident that they can be solved either by existing techniques or by techniques currently under development.”

1) *Ensuring that the training data are sufficiently complete and accurate: ML is of value in applications where requirements cannot be fully specified using representations such as truth tables and Boolean logic. The training data has to include enough cases to make it sufficiently unlikely that a situation not represented in the training data will actually occur in practice. The effectiveness of the training data depends on the statistical correlation between the data and reality.*

2) *Verifying the trained software: because the structure of the trained software is not visible/analysable it is not possible to know what coverage is achieved by testing, and it is also not possible to supplement testing by other means such as static analysis.*

3) *Validation of approximated functionality: the chief advantage of most of the ML methods lies in the generalization of their intended functions into areas of application not specifically defined in beforehand. Precisely this advantage turns into a challenge when it comes to the aspect of validation, namely the check if the final functionality - which has been statistically approximated by the Machine Learning algorithm - “has learned the correct function”.*

4) *Adversarial attacks and missing causality: so-called “adversarial attacks”, e.g. small deviations in the picture to be recognized by a ML algorithm where even small changes in the picture can completely change the categorization of the object. This is among others due to the fact that a ML algorithm does not provide causal relations between input and output, but merely statistical correlations.*

So Artificial intelligence and Machine Learning are still Not Recommended in software architecture for SIL1 to SIL4, and can be used only for basic integrity

TECHNIQUE/MEASURE	Ref	Basic Integrity	SIL 1	SIL 2	SIL 3	SIL 4
13. Artificial Intelligence and Machine Learning	C.3	-	NR	NR	NR	NR

Figure 17: citation from EN 50716 (Table A.3 – Software architecture (7.3))

4.2 REQUIREMENTS FOR THE PERCEPTION SYSTEM

4.2.1 Functional Requirements

The Requirements are based on the requirements for the perception system as defined in FP2-R2DATO D6.3[17].

The perception system is designed to ensure safe and reliable train operations under automated and autonomous conditions by monitoring, detecting, classifying, and interpreting the railway environment. Functional requirements cover a wide set of capabilities, which form the basis for certification:

- **Environment Monitoring and Hazard Detection:** Continuous monitoring of the ego track and its surroundings to detect obstacles, humans, animals, railway agents, and other rolling stock (e.g. *“Any object having an estimated aerea of impactlarger than 0,25 m²shall be detected and considered as an obstacle” Monitor physical environment, Get list of obstacles*).
- **Object Classification and Tracking:** Detected objects must be classified and tracked with unique IDs and trajectories, ensuring that their relevance for train safety can be assessed.
- **Distance and Speed Calculation:** Real-time computation of the distance to obstacles and the relative speed of objects is required for collision avoidance and braking curve adaptation.
- **Environmental Condition Assessment:** The system shall detect and classify adverse conditions such as fog, snow, flooding, fire, or catenary damage, and provide the results as operational inputs.
- **Track Awareness:** Functions include ego track detection and consolidation of the ego-route, enabling correct correlation between detected objects and the train’s journey profile.
- **Safety Envelope Computation:** Calculation of the track-ahead free distance and determination of safe areas form the technical basis for safe decision-making.
- **Self-Monitoring and Evidence Recording:** Internal monitoring ensures availability and consistency, while evidence recording provides traceability for certification and investigation.
- **Signal reading (optional functionality):** The system shall react and interpret lineside signals, during transition between ETCS fitted areas and lineside signaling equipped areas and react non-permanent lineside signals.

4.2.2 Non-Functional requirements – Real-Time Safety

Beyond functional detection and classification, certification requires that the perception system guarantees **real-time performance and operational safety**:

- **Reaction Time**: Sensor data must be processed and obstacle lists updated at configurable (see FP2-R2DATO D6.3[17] requirement ID36.2), ensuring detection and reaction times comparable to or faster than human operators.
- **Trustworthiness and Confidence Metrics**: Outputs shall include confidence levels to allow higher-level modules to adapt their responses.
- **Robustness under Environmental Stress**: The perception system must maintain operational reliability in degraded visibility, adverse weather, and complex operational contexts. Any reduction in performance shall be detected and reported to enable safe fallback.
- **Data Recording and Evidence Provision**: Storage of raw and processed data must be ensured, in respect to the GDPR, for validation purposes and post-incident analysis.
- **Interoperability and Interfaces**: Standardized interfaces with APM, ATP, ADM, REP, and TAS must be provided to guarantee consistent integration into the overall CCS/ATO system.

4.3 ARCHITECTURE OF THE PERCEPTION SYSTEM

4.3.1 General approach to the architecture

The perception architecture has been defined in D6.3 – Architecture and System Specification for Perception System[17] by consolidating WP5 use cases with previous frameworks (X2Rail-4, TAURO), focusing on modularity and adaptability. The architecture follows a **functional layering** approach inspired by the JDL Data Fusion Model, adapted to railway applications:

- **Level 0 – Signal / Feature Assessment**: Sensor input capture and preprocessing (camera, radar, LIDAR, infrared, vibration, weather sensors).
- **Level 1 – Object Refinement**: Classification and localization of detected objects.
- **Level 2 – Situation Assessment**: Analysis of object relationships and dynamics in relation to the ego-track.
- **Level 3 – Impact Assessment**: Evaluation of potential hazards and consequences such as collisions, flooding, or adhesion loss.
- **Level 4 – Performance Monitoring**: Self-assessment of perception system availability and consistency.

This layered approach provides modularity, traceability to requirements, and a clear basis for certification.

4.3.2 High level description of the architecture

At high level, the perception system architecture includes the following stages:

1. Data Acquisition Layer

- Collects raw signals from multiple heterogeneous sensors.
- Receives digital map and localization data for contextual awareness.

2. Sensor Fusion and Object Detection

- Combines sensor inputs using early, mid or late fusion strategies.
- Produces a consolidated list of detected objects including class, position, and motion attributes.

3. Environment and Situation Analysis

- Computes track-ahead free distance and safe areas.
- Monitors areas of interest and evaluates environmental and infrastructure conditions.

4. Decision Support Outputs

- Provides obstacle and environment data, enriched with confidence levels, to APM and ADM.
- Supplies operationally relevant information such as context determination for adhesion estimation (leafes on the track, dry or wet rails), flooding or catenary warnings, and impact detection.
- Consolidates ego-route alignment with train trajectory.

5. Monitoring and Recording

- Performs self-diagnostics and health monitoring.
- Records raw and processed data for traceability and certification evidence.

This high-level design ensures that perception functions can be validated systematically against defined requirements, while leaving implementation flexibility to system suppliers. It forms the foundation for certification and integration into the broader CCS/ATO framework.

4.4 SAFETY REQUIREMENTS

From September 2024 until the end of July 2025, FP2-R2DATO WP8 Task 2 conducted the safety analysis of the Perception system (PER). Under the lead of FT, SBB, DB, CAF, Alstom, SNCF, Siemens, NRD, SMO, RENFE, and GTSD participated in this analysis.

The PER safety analysis identifies the needed requirements to ensure an adequate level of safety for PER considering its use under ETCS Full Supervision. It is mainly based on PER Use Cases derived from FP2 R2DATO Deliverable 5.2 "Documentation of Use Cases for Perception system" D5.2 revision 05 and classified into 5 main categories:

1. Object Detection
2. Signal Reading
3. Surveillance of Location Associated with Increased Risk of Collision
4. Detection of Infrastructure Defects and Anomaly Detection
5. Automatic joining of multiple units

The PER system has the task of detecting and reporting anomalies and performs continuous monitoring of the anomalies in real time. Therefore, WP8 Task 2 safety analysis has considered PER system to be operating in “continuous mode” according to

the models recognized by the safety standards considering that PER detection is solicited more than once per year.

The Safety analysis approach followed a top-down approach only. In a first step, it has been conducted a qualitative estimation of the risk caused by the anomalies, based on how often the anomalies are expected to occur and on their associated credible accidents. External safety barriers independent of PER have been considered in this qualitative estimation. In a second step, if the risk without PER is not already acceptable, quantitative safety requirements for PER functions (TFFR and SIL) have been defined while assumptions and SRAC (safety related application conditions) have been listed.

Based on WP8 Task 2, for the use cases that have been analyzed and under the assumptions taken, PER system needs to perform some safety relevant functions, **up to SIL2** for the Use Cases which have been analyzed.

SIL2 statement is valid only under ETCS Full Supervision or with an active ATP. Without an active ATP, for lateral signal reading Use Case, PER system must be SIL4.

The safety analysis could not be conclusive for certain scenarios that still need to be clarified.

The list of not-completely-analyzed Use Cases for which the conclusion of SIL2 requested is not valid, is as follows:

N° Use Case in D5.2	Name of the Use Case	Required complement
UC5.2-0008	Rail Adhesion Estimation using Perception	Not defined as the level of safety requirements is to be defined according to the BAMS system performances. This scenario is not relevant for a quantified assessment of the safety level.
UC5.2-0016	Handle reduced visibility and detection capabilities	PER operating range needs to be clarified: The PER module must be compatible with current operating conditions (365 Day/24 hours + environmental conditions): be independent of environmental conditions

		(measurement principle, technology, processing, etc.) or be able to detect when it leaves the correct operating range (faults, saturation, etc.) in order to reduce train speed and raise an alarm.
UC5.2-0024	Impassable Broken or Buckled Rail	It needs to be specified what is required of perception functions on the individual train. Without this information, it is not possible to allocate TFFR/SIL for these functions.
UC5.2-0028	Automatic joining of electric and/or diesel multiple units	Conditions before joining aren't defined in the 5.2 document. Not possible to know how if PER is used here.
UC5.2-0041	Evaluating ego track	Not defined Constraints to be exported: 1- Further work for coordination and integration between Perception and ASTP is needed 2- A wrong identification of the ego-route and determination of track-ahead free distance could lead to a catastrophic accident without a driver in the train cabin (e.g., collisions with obstacles within the clearance gauge, collisions with other vehicles in complex environments such as entrance track to stations, etc.). 3- Up to GoA2 the train driver is a relevant barrier to compensate for unprecise

		<p>ego-route determinations that could lead to collisions in complex environments, as well as to determine track ahead free distances (e.g., detection of obstacles invading the clearance gauge).</p> <p>4- The SIL requirements for determination of the ego-route are conditioned by the physical environment. E.g., in metro systems with Platform Screen Doors (PSDs) running through tunnels, a faulty determination of the ego-route won't be as important as in mainline systems, since the metro infrastructure is designed to avoid trespassing, presence of unexpected obstacles, etc.</p>
UC5.2-0055	Transition from/to ATP Lineside signalling area / Class A area	Depends on Safety Level of ASTP required as an external technical safety barrier.
UC5.2-0058	Driving according to journey based on lateral signalling	Depends on Safety Level of ASTP required as an external technical safety barrier.

In a next step, this safety analysis needs to be reviewed by the System Requirement Specifications group in order to update the existing Deliverable FP2-R2DATO 6.3 (i.e. “Architecture and System Specification for Perception System” D6.3)[17].

5 CERTIFICATION PROCESS DESCRIPTION

This chapter provides a comprehensive description of the certification process for perception-based functions (PER) on rolling stock within the R2DATO WP 6.4 context. It refers to the Technical Specification for Interoperability for locomotives and passenger rolling stock (TSI LOC&PAS)[4]. The actors mentioned in the TSI LOC&PAS are used in the process description. The document is based on the requirements specific to perception systems, and proposes a workflow with clear responsibility assignments. The system under concern in this chapter is the perception system. The viewpoint is what additional aspects must be handled if a train is equipped with a perception system. The safety analysis is made for the train and dependent from the chosen architecture the perception functions must provide a defined integrity.

The process aligns with the requirements and guidance presented in Chapter 5 of this deliverable, particularly the consolidated TSI LOC&PAS and associated TSIs. The chapter serves as a blueprint for railway undertakings (operators), infrastructure managers and vehicle suppliers to navigate conformity assessment, safety assurance and authorization procedures when deploying AI-enabled perception functions in the European railway system.

5.1 ACTORS IN FORCE

Ensuring safe and interoperable deployment of perception-based functions requires clear definition of the actors and their respective responsibilities. In Europe, the regulatory framework relies on a combination of European and national bodies, as well as industrial stakeholders. This section describes each actor's role in the certification and authorization process, emphasizing their interactions and accountability.

5.1.1 European and national authorities

European Union Agency for Railways (ERA) – ERA oversees the application and evolution of the Technical Specifications for Interoperability (TSIs), which define the technical and operational standards necessary to meet essential requirements and ensure interoperability of the EU railway system era.europa.eu. In the context of vehicle authorisation, ERA operates the OneStopShop (OSS) platform and may act as the authorising body for vehicle types intended for crossborder operation. ERA provides opinions and technical advices on TSI interpretations and coordinates the work of national authorities.

National Safety Authorities (NSAs) – NSAs are responsible for safety certification and authorisation of railway undertakings and infrastructure managers in their respective Member States. For vehicle authorisation limited to national networks, the NSA acts as the

authorising body. It ensures that the applicant's safety case meets national rules and EU requirements, including TSIs and the Common Safety Methods (CSMs).

Notified Bodies (NoBo) – **NoBos** perform independent conformity assessment against TSIs. For rolling stock, this includes verifying that vehicles and their constituents meet the requirements of the TSI LOC&PAS, TSI CCS, TSI OPE and other applicable TSIs[4] era.europa.eu. In cases where the perception system or its integration is considered an innovative solution under the TSI, the **NoBo** collaborates with ERA and the European Commission to agree on assessment principles.

Designated Bodies (DeBo) – Designated bodies perform conformity assessment against national technical rules that complement the TSIs. They are involved when a Member State has notified specific rules not covered by the TSIs.

Assessment Bodies (AsBo) – Under the Common Safety Method for Risk Evaluation and Assessment (CSM RA), an independent assessment is required for significant changes. **AsBos** evaluate the safety justification, including hazard identification, risk acceptance principles and adequacy of risk control measures. They provide an Assessment Report that is a key element of the safety case.

5.1.2 Industrial stakeholders

Railway Undertaking (RU) – Operator – The operator procures and operates rolling stock for specific services and routes. In the context of perception systems, the RU is accountable for defining the Operational Design Domain (ODD) for the intended deployment, ensuring compatibility of the vehicle with the route and operational conditions, and integrating perception-driven outputs into operational procedures (e.g., driver information, remote supervision). The RU should be accountable for the ODD, because it defines together with the IM the operational context of the system including the environmental conditions under which the operation must be possible.¹ The RU coordinates with the infrastructure manager and supplier to obtain necessary data (HD maps, simulation inputs) and ensures that ODD boundaries are respected during operations and updates. To reduce the complexity of the ODD definition it would be beneficial if a general or easy adaptable standard ODD definition exists.

Infrastructure Manager (IM) – The IM owns and maintains the railway infrastructure, provides access to the network and supplies infrastructure data. For perception-based functions, the IM is responsible for delivering an authoritative high-definition (HD) map of the route, including track geometry, signaling, speed restrictions and relevant environmental features. The IM also provides inputs required for routes-specific simulation, such as topography, gradients, line speeds and lineside asset characteristics. It

¹ If the Supplier would define the ODD it might not match the operational context and the conditions under which the operation must be possible. This may lead to a lot of trouble during project execution. Additionally, there would be no level playing field, because the suppliers might offer solutions for different ODDs that might not be compared during the bid phase. Suppliers might reduce the ODD and such the costs for the development of the system.

collaborates with the RU on ODD definition by supplying environmental and infrastructure constraints.

Vehicle Supplier / Manufacturer (Supplier) – The supplier designs and manufactures rolling stock, integrates perception systems and related automation functions, and demonstrates compliance with general TSI requirements. When a perception system is considered an innovative solution, the supplier must propose an assessment approach and provide evidence of safety and performance. Responsibilities include developing and maintaining the PER-grade HD map representation, providing vehicle and perception models for simulation, defining the product ODD (capability envelope), and ensuring that the perception function interfaces properly with train control and management systems.

Perception System Supplier (if distinct) – In cases where the perception module is provided by a specialist supplier, this entity is responsible for the design, training and verification of the perception algorithms, sensor integration, and provision of performance metrics and hazard analyses. The PER supplier collaborates with the vehicle supplier to integrate the system into the vehicle architecture.

5.1.3 Interaction overview

The certification of perception systems requires coordinated interactions between the described actors. This chapter illustrates the overall relationship between the stakeholders, highlighting that the **RU**, **IM** and **Supplier** form the primary stakeholder triangle responsible for preparing the data and evidence package. They interface with the **NoBo** and **DeBo** for conformity assessment and with the **AsBo** for safety assessment. **ERA** and **NSAs** provide regulatory oversight and authorization decisions.

The proposed responsibilities are intended to serve as a basis for discussing this complex topic, as stakeholders possess varying competencies in different countries and projects. Defined default responsibilities support easier project setup, assure building up the right competences, tools and processes at the stakeholders and reduce tensions and costs in the projects.

The relationship between RU and IM might be challenging, since the implementation of driverless trains reduces the number of TFs needed on board thus being beneficial for the RU to offer their service, e.g. in order to answer challenges due to TF shortage while the IM is required to provide additional assets beyond their current scope of responsibility without having a direct benefit for the RU.

Another area to address concerns is the provision of the simulation environment and standards how the information about physical objects must be provided to support the simulation, ensuring simulations with an appropriate small reality gap for training and validation purposes. The quality of object models within the ODD, including their physical parameters, may depend on how suppliers design, manage, and verify solutions.

Below is a summary of potential interaction:

1. **RU/IM ↔ Supplier:** The **RU** and **IM** define operational requirements as well as ODD boundaries and content; the Supplier delivers the vehicle and perception system capable of meeting the requirements inside the ODD.
2. **RU ↔ IM:** The **RU** requests HD map data and route simulation inputs; the **IM** supplies authoritative infrastructure data and informs about planned changes. The **RU** verifies map accuracy and simulation fidelity for operational readiness. The **IM** provides the HD map for the project and is responsible for the maintenance.
3. **Supplier ↔ IM:** The potential **Suppliers** provide required specifications extension from the HD map standard including e.g. improved resolution as well as project specific assets for the simulation as part of their bid. These extensions might be important to address special requirements in a project or may help to reduce the overall costs of the solution offered. The offer from the supplier should / may include services that provide the information for the HD-map extensions. The **IM** and **RU** supply the HD-Map data accordingly during the project phase.
4. **RU/Supplier/IM ↔ NoBo/DeBo:** Conformity assessment bodies verify compliance with TSIs and national rules based on the dossier provided. Additional evaluation is required for innovative aspects like HD maps and simulation.
5. **RU/Supplier/IM ↔ AsBo:** The stakeholders submit the hazard analysis and safety argument to the **AsBo**, who independently assesses the risk management and issues an Assessment Report.
6. **ERA/NSA ↔ RU:** The **RU** applies for authorization to place the vehicle in service; **ERA** or the national **NSA** evaluates the dossier (including **NoBo/DeBo** certificates and **AsBo** reports) and issues the authorization after the assessment approach and relevant criteria for safety and performance have been agreed upon with ERA/ the N

5.2 GENERAL CERTIFICATION WORKFLOW OVERVIEW

The general workflow for certifying perception-equipped rolling stock is designed to fulfill the TSI LOC&PAS[4] and its associated conformity assessment modules. It encompasses definition of requirements, data preparation, design and verification, safety assessment and authorization.

5.2.1 Concept and scoping

The process begins when an RU decides to procure a vehicle equipped with a perception system for a specific line. At this stage, the **RU** formulates the **service concept**, identifies the intended **Operational Design Domain (ODD)** using for example the taxonomy referenced in Chapter 4 (e.g., DIN SPEC 99004) and ensures alignment with the use cases and requirements consolidated in WP 6.3.

5.2.2 Requirements and architecture alignment

In this phase of an instance of GoA3/4 project, the **RU**, **IM** (optionally) refine together with the **Supplier** the functional and nonfunctional requirements of the perception system based on the existing standards. Requirements include detection performance thresholds, latency limits, availability, failsafe behavior, integration with APM (Automatic Processing Module), and environmental constraints. The architecture from WP6.3 and the hazard

analysis from WP8.2 inform the design. The required subset of the requirements from WP 6.3 is defined and additional requirements are provided to the potential suppliers.

5.2.3 Data foundations

Reliable data is essential for design, simulation and validation. The **IM** provides the infrastructure information required for the **HD-map to the RU** containing track geometry, gradients, curvature, signaling locations, speed limits, platforms and other line-side assets. The **RU** transforms this information into a PER-grade representation according to the standards for a HD-map. The supplier may provide extensions to enable the fulfillment of the project requirements. The **RU** establishes acceptance criteria for the HD-Map and manages change control. In parallel, the **IM** and **RU** provide the **route simulation dataset** covering infrastructure parameters, operational scenarios (normal and degraded), environmental conditions and communications coverage (e.g. containing parts of the route where a remote operator may take over control). The Supplier delivers vehicle dynamics models, sensor and perception models, and defines scenarios to simulate fault injection and edge cases. All datasets are version controlled and traceable to requirements and ODD assumptions.

5.2.4 Safety and risk management

The Common Safety Method for Risk Evaluation and Assessment (CSM RA) governs safety assurance. Especially in the first projects, when the stakeholders have no comprehensive experience, it is helpful to perform the risk assessment together to have access to the information and experience of all stakeholders to avoid risk over or underassessment as well to avoid missing risks. The stakeholders jointly perform a hazard identification exercise, deriving safety requirements and risk control measures. The level of risk acceptance is determined according to the CSM risk acceptance principles. A **hazard log** is established and maintained. Because perception systems may rely also AI, the supplier project must consider AI-specific hazards (as e.g. described in the LAISC methodology, Chapter 4) and allocate additional mitigation measures. In this activity the brake down of the results of the safety analysis to the perception system based on the system architecture is done by the supplier. The AsBo is engaged early to review the methodology.

5.2.5 Verification and validation strategy

Verification and validation activities cover unit testing, integration testing, simulation, hardware-in-the-loop, software-in-the-loop and on-track tests. Test cases are derived from requirements and hazards and are mapped to ODD scenarios. It must be assured that tests cover the ODD according to the expected environmental conditions during the planned operation. The **virtual test field** (Chapter 3.8.5.3) is employed to perform simulation and generate evidence for detection performance, robustness, latency and integration. Interface tests ensure correct data exchange between perception functions and APM. Test results are compared against acceptance criteria and used to refine models and the ODD definition. All evidence is documented and traceable.

5.2.6 Conformity assessment

Upon completion of design and testing, a conformity assessment is conducted. The Supplier prepares a **Technical File** summarizing compliance with the TSI LOC&PAS (and other TSIs, such as CCS or OPE if relevant). The NoBo reviews the Technical File, performs necessary inspections (e.g., design review, test witnessing, manufacturing surveillance) and provides a conformity assessment. Where national rules apply, the DeBo conducts a similar assessment.

5.2.7 Safety assessment

The RU compiles a **Safety Case** for the train including the PER system including the hazard log, safety requirements, risk analyses, verification evidence and independent assessment reports typically supported by the **Supplier**. The AsBo evaluates risk management and verifies that the risk control measures are adequate and correctly implemented. An **Assessment Report** for the train offered by the supplier including the PER system is issued and becomes part of the authorization dossier.

5.2.8 Authorization and deployment

Finally, the **RU** applies for authorization to place the vehicle in service using **ERA's** OneStopShop or the relevant **NSA**. The application includes the EC Certificate, the Safety Case, the AsBo report and supporting documentation. **ERA** or the **NSA** verifies the completeness of the dossier and issues an authorization decision. The **RU** then deploys the vehicle on the specified route. The process includes monitoring and maintenance obligations; relevant modifications to the perception system, HD-map or ODD trigger reassessment under the CSM RA.

5.3 SPECIAL ASPECTS AND REQUIREMENTS FOR PERCEPTION SYSTEMS FOR THE CERTIFICATION

While the general workflow aligns with existing rolling stock certification practice, perception systems introduce specific challenges. AI-based perception functions rely on complex algorithms, diverse data sources and dynamic operational contexts. This section details PER-specific certification aspects and requirements that must be addressed to satisfy conformity to norms and standards and safety assurance. In chapter 4 methods that may help during the certification process have been compiled for example in the DEEL[1], safe.trAIIn[11] and Confiance. AI project. Still additional research is required to cover all aspects of the certification process.

Due to the high challenges from the topic and the potential wide field of application projects up to now no consensus is achieved on general aspects. One example is, if the demand mode described in IEC 61508 [7](see chapter 3.1) be applicable for the safety case. The safety assessment must be made for every project (like S-Bahn München or commuter rail in the Ile de France area) separately because too many aspects may be different and must be considered. The certificate may consist of general aspects like the

used processes by the supplier that are used for different projects. Other aspects of the certificate will be project specific.

5.3.1 Modular operational design domain (ODD)

The introduction of a modular ODD approach offers significant advantages in the certification of perception systems. By dividing the overall system into clearly defined, functional subdomains (modules), specific operational conditions and requirements can be described, assessed, and certified separately. This enables a more precise and traceable definition of the perception system's operational boundaries, both in terms of environmental conditions and infrastructural or operational scenarios.

A modular ODD facilitates the reuse of already certified modules across different projects, thereby reducing the assessment effort required for system adaptations. Modifications to the perception system or infrastructure can be evaluated in a targeted and efficient manner, as only the relevant modules need to undergo reassessment. This results in increased flexibility for further development and adaptation of systems to new areas of application or changing requirements.

Furthermore, the modular structure supports traceability of requirements, tests, and evidence within the certification process. The ability to directly relate safety evidence to individual ODD modules helps reduce complexity and structure the documentation. This is particularly relevant when multiple suppliers or different technologies are integrated within a complete system.

- Person module: Definition of persons and their possible poses within the ODD, such as standing, walking, or lying down individuals.
- Vehicle type module: Description of the different types of vehicles operating on the track, for example, locomotives, freight cars, road vehicles at level crossings, or service vehicles.
- Track type module: Modules for various types of track infrastructure, such as tracks with special signaling, level crossings, tunnels, or sections with differing infrastructure characteristics.

These modules allow requirements, testing, and safety evidence to be specifically assigned to individual subdomains.

In summary, a modular ODD approach contributes to a more efficient, transparent, and scalable certification of perception systems. It allows for focused risk assessment tailored to specific operational domains and promotes innovation through the reuse and flexible adaptation of certified modules.

5.3.2 Performance and acceptance metrics

Perception performance must be quantified and linked to safety targets. Key metrics include:

- **Detection probability (recall) and false alarm rate** – The system must detect relevant objects (e.g., obstacles, persons, animals) with a probability meeting or exceeding the specified threshold. As discussed in Chapter 4, the acceptance criterion may be expressed for example in terms of Probability of Failure on Demand (PFD) equal to 1 minus recall.
- **Latency** – The time between sensor observation and decision output must not exceed the limit defined by the operational scenario, ensuring the train can break or react appropriately.
- **Availability and fault tolerance** – The perception system must remain available during the mission with a defined mean time between hazardous failures. Redundancy, monitoring and fail-safe modes contribute to availability. The perception system should also indicate degraded performance due to adverse weather conditions.
- **Robustness to environmental conditions** – Performance must be maintained across the ODD's environmental range (weather, lighting, speed, track conditions) as specified by the ODD definition.
- **Uncertainty estimation** – The system must provide a measure of detection confidence or uncertainty, enabling monitors to decide when to rely on results.

Performance metrics must be verified through simulation, laboratory and on-track tests. Acceptance thresholds are documented in the requirements specification and validated through test campaigns.

5.3.3 Dataset and labelling governance

AI perception models depend on high-quality training, validation and test datasets. Certification requires evidence that datasets:

- **Cover the ODD** – Data collection must represent the range of scenarios, environments and objects specified in the ODD. This includes diverse weather, lighting, speeds and obstacle types.
- **Define Scenario catalogue** - These scenarios support the quality assurance in rare critical situations like „Person approaching the ego route“ or „Hidden Person jumps on tracks“. These scenarios should be tested for risk reduction under relevant ODD conditions.
- **Ensure traceable provenance and version control** – Each dataset element must be traceable to its origin (real sensor data or synthetic generation) and labelled version. Changes to datasets must be controlled, with retraining and revalidation performed when necessary.
- **Use consistent annotation guidelines** – Labelling must follow agreed guidelines to reduce ambiguity and bias. The safe.trAI[n] methodology emphasizes clear labelling processes and monitoring of labelling accuracy.
- **Address ethical and privacy considerations** – Data used for training must comply with GDPR and other privacy regulations. Anonymization or removal of personal data is required where applicable. To date, there is no known anonymization method for person detection available that can reliably prevent artifacts which may compromise detection accuracy and the validity of results.

The data governance plan forms part of the assurance dossier and supports both safety assessment and conformity assessment.

5.3.4 HD map quality and change management

The HD map as defined in ERJU R2DATO D27.2 – [19] **Specification** of the **Digital Register** implementation is a core data asset for perception localization and situational awareness. Certification requires the following:

- **Positional accuracy** – The map must provide position information with a precision adequate for perception and train control functions. Tolerances for absolute and relative positioning must be defined and verified.
- **Topology and attribute completeness** – All relevant track elements (curves, gradients, junctions), signals, speed restrictions and assets must be included. Missing elements or outdated information creates potentially safety hazards.
- **Versioning and updates** – The HD map must be version controlled, with a change management process that defines how updates are validated and rolled out. The **RU** and **IM** agree on triggers for revalidation (e.g., infrastructure changes, detection of map errors).
- **Data format and interfaces** – The map must be provided in a format compatible with the perception system and train control architecture. It should support dynamic loading and align with vehicle localization coordinates.

Evidence for map quality includes survey reports, accuracy tests and change logs. The map specification and quality report are part of the Technical File for conformity assessment.

5.3.5 Simulation fidelity

Simulation plays a major role in verification and validation. Requirements include:

- **Infrastructure model fidelity** – The simulated track geometry, gradients, curvature and signaling logic must accurately reflect the real route. The **IM** provides authoritative data; the Supplier implements high-fidelity models.
- **Vehicle and sensor models** – The simulation must incorporate realistic vehicle dynamics, sensor characteristics (field of view, noise, failure modes) and perception algorithms. Model parameters must be validated against real-world measurements.
- **Scenario coverage** – The scenario library must cover the ODD, including nominal and degraded operations, rare events and edge cases. Scenarios are defined jointly by the **RU** (operations), **Supplier** (fault injection) and **IM** (infrastructure constraints).
- **Correlation and representativeness** – Simulation results must be correlated with on-track test data to validate model accuracy. Differences are analyzed and models updated accordingly.

Simulation evidence includes configuration files, model descriptions, scenario definitions and correlation reports. The results form part of the verification evidence submitted to the **NoBo** and **AsBo**.

5.3.6 Interfaces and interoperability

Perception systems must interoperate with the train control and communication subsystems:

- **APM integration** – Perception outputs (e.g., obstacle detections, warnings) must be communicated to the **APM** Automatic Processing Module onboard.
- **Failsafe fallbacks** – In case of perception malfunction or uncertainty, the system must revert to safe states (e.g., reduced speed, emergency braking) and maintain safe operations. Integration with existing safety functions (e.g., train protection) is critical. The safe state may depend on the specific project and might also include the "safe take over from a remote train driver".
- **Failure mode reporting** – The system must provide failure mode information and health status to enable monitoring by the **RU** and maintenance personnel.

5.3.7 Safety case evidence map

For perception systems, the safety case must trace hazards to requirements, tests and evidence. A typical evidence map includes:

- **Hazard Log** – Listing AI-specific and system hazards, their classification and risk mitigations.
- **Requirements Traceability Matrix** – Mapping hazards to safety requirements and design features.
- **Verification and Validation Records** – Test specifications, results and analyses demonstrating that requirements are met.
- **Independent Assessment Reports** – **AsBo** assessment reports and **NoBo/DeBo** certificates.
- **Operational Monitoring Plan** – Procedures for runtime monitoring, performance tracking, ODD compliance and re-assessment triggers.

The evidence map ensures that all aspects of the perception system's design and operation are covered and provides transparency for assessors and authorizing bodies.

5.4 EXTENDED WORKFLOW OVERVIEW

The extended workflow translates the general process into a detailed, phased plan with clear roles, responsibilities and deliverables. It emphasizes collaboration between **IM**, **RU** and **Supplier** while incorporating interactions with assessment bodies and authorizing entities. Each stage includes inputs, activities, outputs and gate criteria. RACI tables summarize responsibilities for key data assets.

5.4.1 Stage 0 – Initiation and governance

Purpose: Establish project organization, governance and documentation control. Define team roles and processes for requirement management, change control and configuration management.

Inputs: Contractual agreements between RU, **IM** and **Supplier**; project charter; regulatory framework (TSIs, CSM RA, national rules). Previous deliverables from WP5 (use cases and requirements) and WP6 (architecture).

Activities:

- Set up governance structure and assign responsibilities.
- Define document management procedures and configuration control.
- Establish regulatory baseline (TSI LOC&PAS applicability, other TSIs, national rules).
- Plan stakeholder engagement (**NoBo**, **DeBo**, **AsBo**, **ERA/NSAs**).

Outputs: Governance plan; document control procedure; initial regulatory matrix; stakeholder engagement plan.

Gate criteria: Governance plan approved by **RU**, **IM** and **Supplier**; regulatory baseline verified by legal experts.

5.4.2 Stage 1 – ODD baseline and hazard scenario definition

Purpose:

Define the initial Operational Design Domain (ODD) and the relevant hazard scenarios for the deployment of the perception system, combining product capabilities and route constraints. This includes a systematic identification and documentation of scenarios, especially those at, near or beyond the ODD boundaries.

Inputs:

- **Project ODD:** Provided as part of the requirements by the Railway Undertaking (RU) and Infrastructure Manager (IM), describing the perception system's capability boundaries (e.g., detectable object classes, speed ranges, weather conditions, occlusion handling).
- **Route/environment constraints:** Provided by the IM (infrastructure characteristics, environmental features) and RU (operational procedures, service types).
- **ODD taxonomy and definitions:** As referenced in Chapter 3.

Activities:

- RU, IM, and Supplier conduct workshops to align their understanding of the ODD.

- Identify and define concrete scenarios, with a particular focus on those at the edge or outside the capability envelope of the perception system.
- For scenarios outside the ODD, define mitigations (e.g., speed restrictions, fallback procedures).
- Document ODD attributes (environmental, temporal, spatial, operational) and produce both machine-readable and human-readable ODD specifications.
- Explicitly document the scenario catalogue, including boundary and out-of-ODD scenarios, as part of the ODD specification.

Outputs:

- ODD specification
- Structured scenario catalogue, with special attention to boundary and out-of-ODD scenarios
- ODD conformance matrix
- Initial gap analysis and mitigation plan
- Documented list of mitigations for scenarios outside the ODD

Gate Criteria:

- ODD specification and the scenario catalogue (with boundary and out-of-ODD scenarios), is signed off by the RU (accountable), with IM and Supplier approval.
- All identified scenarios and corresponding mitigations are documented and agreed.
- ODD and scenario catalogue are integrated into the requirements management system.

5.4.3 Stage 2 – HD map definition

Purpose: Prepare a high definition- infrastructure map supporting perception localization, scene understanding and simulation.

Inputs:

- Authoritative infrastructure data from the **IM**, including geometry, curvature, gradients, signaling, speed restrictions, stations, level crossings and other assets.
- Map specification (format, resolution) defined by the R2DATO WP27.
- Acceptance criteria from the **RU** (accuracy thresholds, completeness requirements, update frequency).

Activities:

- **IM** collects and validates raw infrastructure data, ensuring that information is accurate and up to date. **RU** transforms the raw data into a PER-grade HD map, defining layers (lane center lines, boundaries, landmarks, signal positions) and coordinate frames aligned with vehicle localization according to the specification by R2DATO WP 27.

- **RU** reviews the map against operational expectations, performs consistency checks and defines change control procedures.

Outputs: PER-grade HD map; map quality report; change management plan; map version control records.

Gate criteria: Map accepted by **RU**; quality metrics meet specified tolerances; change management process established.

5.4.4 Stage 3 – Simulation environment preparation

Purpose: Establish a digital twin environment for design verification and validation.

Inputs:

- Infrastructure model from the **IM**; route scenarios and operational constraints from the **RU**.
- Models of Objects in the ODD provided by the **IM** to gain synergy between multiple projects in the same region. The IM is the only stakeholder with typically a strong regional reference
- Vehicle dynamics model and perception/sensor models from the **Supplier**.
- ODD specification; HD map from **IM** and **RU**.

Activities:

- Supplier integrates infrastructure and vehicle models into a simulation platform; implements sensor models and perception algorithms.
- **RU** defines operational scenarios (normal timetable, degraded modes, yard operations) and edge cases; **Supplier** implements scenario definitions and fault injection.
- **IM** provides environmental parameters (e.g., communications coverage, occlusion zones) and verifies infrastructure model fidelity.
- Define correlation plan to validate simulation results against future field tests.

Outputs: Simulation configuration baseline; scenario library; correlation plan; simulation readiness report.

Gate criteria: Simulation environment validated; scenario library covers the ODD; correlation plan approved by **RU** and **IM**; simulation ready for use in verification phase.

5.4.5 Stage 4 – Safety, hazard analysis and architecture

Purpose: Perform the CSM RA process, identifying hazards, allocating safety requirements and planning evidence generation.

Inputs: ODD specification; architecture and requirements; HD map and simulation artefacts; AI safety concerns list (LAISC) from Chapter 4.

Activities:

- Conduct hazard identification sessions; include AI-specific concerns (e.g., inaccurate data labels, model drift) and conventional system hazards.
- Apply risk acceptance principles (codes of practice, reference values, explicit risk estimation) to determine acceptable risk levels.
- Define system architecture
- Allocate safety requirements to system components, interfaces and operational procedures.
- Plan verification activities (simulation tests, lab tests, field tests) to collect evidence.
- Engage **AsBo** for preliminary assessment of hazard analysis and plan.

Outputs: Hazard log; safety requirements specification; verification plan; preliminary **AsBo** feedback.

Gate criteria: Hazard log reviewed and approved; safety requirements allocated; verification plan accepted by **RU**, **IM** and **Supplier**; **AsBo** indicates readiness for next stage.

5.4.6 Stage 5 – Verification (lab and simulation tests)

Purpose: Execute laboratory and simulation tests to verify compliance with requirements.

Inputs: Verification plan; simulation environment; perception system prototypes; HD map and datasets; ODD specification.

Activities:

- Unit tests on perception algorithms using labelled datasets, verifying detection performance, robustness and latency.
- Integration tests to verify sensor fusion, perception output to **IM** interfaces and fail-safe behavior.
- Simulation tests and real time data covering all ODD scenarios, including rare events and edge cases; collect metrics and identify performance gaps.
- Update hazard log and ODD specification based on test results.

Outputs: Test reports with metrics; issues and bug reports; updated models and requirements; evidence for conformity assessment.

Gate criteria: Performance metrics meet acceptance thresholds; integration verified; simulation coverage adequate; readiness for field testing confirmed.

5.4.7 Stage 6 – Field validation

Purpose: Validate system performance and safety under real-world conditions on the target route.

Inputs: Verified prototypes; HD map; operational scenarios; test trains; authorizations for testing from IM and NSA.

Activities:

- Conduct field tests on the designated route within the ODD; collect perception outputs, vehicle behavior and environmental data.
- Validate detection performance, latency and robustness in real conditions; compare with simulation results and update models.
- Test integration with train control systems and operational procedures; verify fail-safe transitions and driver/operator interactions.
- Monitor compliance with safety requirements and update hazard log.

Outputs: Field test reports; correlation analyses; updated models and ODD specification; finalized verification evidence.

Gate criteria: Field test results meet performance and safety requirements; correlation between simulation and field data validated; no unresolved critical issues; ready for conformity assessment.

5.4.8 Stage 7 – Conformity and safety assessment

Purpose: Obtain official assessments required for authorization.

Inputs: Technical File, HD map and simulation evidence, verification and validation reports, Safety Case, hazard log, ODD specification, **AsBo** preliminary feedback. The CENELEC standards EN 50126, 50128/50657, 50129.

Activities:

- Supplier compiles the Technical File, including design descriptions, compliance analyses and test evidence; submits to **NoBo**.
- **NoBo** reviews documentation and carries out inspections; issues EC Certificate of Verification if compliance with TSI LOC&PAS and other TSIs is proven.
- **DeBo** assesses compliance with national rules if applicable; issues corresponding certificates.
- RU submits the Safety Case to **AsBo** for independent assessment; **AsBo** evaluates risk management and verifies that safety requirements are satisfied; issues Assessment Report.

Outputs: EC Certificate(s) of Verification; **DeBo** certificate(s); AsBo Assessment Report; final Safety Case.

Gate criteria: Certificates and Assessment Reports issued without blocking findings; Safety Case approved; dossier ready for authorization application.

5.4.9 Stage 8 – Authorization and entry into service

Purpose: Obtain authorization to place the vehicle in service and commence operations.

Inputs: Complete authorization dossier (EC Certificates, Safety Case, Assessment Report, ODD specification, HD map and simulation evidence, operational procedures).

Activities:

- **RU** submits authorization application via **ERA's** One-Stop-Shop or to the national **NSA**, depending on operation scope.
- **ERA/NSA** evaluates the dossier; verifies that TSIs and national rules have been met; may request additional information.
- Upon successful evaluation, **ERA/NSA** issues the authorization decision; updates registers (ERATV, RINF) as required.
- **RU** trains staff, implements operational procedures, and deploys the vehicle on the route. Monitoring obligations (e.g., run-time monitoring, ODD compliance) are initiated.

Outputs: Authorization decision; entry-into-service procedures; monitoring plan activation.

Gate criteria: Authorization granted; operational readiness confirmed; monitoring processes in place.

5.4.10 Stage 9 – Change and re-assessment

Purpose: Manage changes during operation and ensure continued compliance.

Inputs: Change requests (e.g., HD map updates due to infrastructure works, perception algorithm updates, ODD modifications); monitoring data; safety reports.

Activities:

- Evaluate change impact; determine whether changes are significant under the CSM RA and TSI rules.
- If significant, perform hazard analysis, update ODD specifications, and develop additional verification activities.
- Submit changes to assessment bodies (NoBo/DeBo, AsBo) as required; obtain revised certificates or Assessment Reports.
- Update authorization dossier and obtain revised authorization if necessary.

Outputs: Change analysis reports; updated hazard log and ODD; revised evidence; updated authorization where applicable.

Gate criteria: Change approved by relevant authorities; updated monitoring plan; assurance that residual risk remains acceptable.

5.4.11 RACI summary tables

The following tables summarize roles and responsibilities using the RACI (Responsible, Accountable, Consulted, Informed) model. Cells contain short labels rather than long sentences.

- HD infrastructure map**

Activity / Asset	R	A	C	I
Master infrastructure data collection	IM	IM	Supplier, RU	NoBo, AsBo
PER-grade HD map generation	RU	RU	IM, Supplier	NoBo, AsBo
Map quality verification and acceptance	RU	RU	IM, Supplier	NoBo, AsBo
Change management and versioning	IM	IM	Supplier, RU	NoBo, AsBo

- Simulation inputs**

Activity / Asset	R	A	C	I
Infrastructure model for simulation	IM	IM	Supplier	NoBo, AsBo
Vehicle and PER models	Supplier	Supplier	RU, IM	NoBo, AsBo
Operational scenario definition	RU	RU	Supplier, IM	NoBo, AsBo
Simulation configuration and campaign integration	RU	RU	Supplier, IM	NoBo, AsBo

- Operational Design Domain (ODD) definition**

Activity / Asset	R	A	C	I
ODD definition	RU	RU	Supplier, IM	NoBo, AsBo
Route/environment ODD definition	IM	RU	Supplier	NoBo, AsBo
Deployment ODD consolidation	RU	RU	Supplier, IM	NoBo, AsBo, ERA/NSA
ODD monitoring and update	RU	RU	IM, Supplier	NoBo, AsBo, ERA/NSA

5.4.12 Milestone and evidence plan

Table 6-1 summarizes the key stages, primary outputs and gate criteria.

Stage	Main Outputs	Gate Criteria
0 – Mobilization	Governance plan; document control procedure; regulatory matrix	Approval by RU, IM, Supplier

Stage	Main Outputs	Gate Criteria
1 – ODD baseline	ODD specification; gap analysis	Signed off by RU; ODD, Supplier integrated into requirements
2 – HD map	PER-grade map; quality report; change plan	Map accepted by Supplier; quality metrics met
3 – Simulation environment	Simulation baseline; scenario library; correlation plan	Simulation validated; scenario coverage adequate
4 – Safety analysis	Hazard log; safety requirements; verification plan	Hazard log approved; AsBo feedback positive
5 – Verification (lab/sim)	Test reports; updated models	Metrics meet thresholds; simulation coverage adequate
6 – Field validation	Field test reports; correlation analyses	Field results satisfactory; ready for assessment
7 – Conformity and safety assessment	EC Certificates; DeBo certificates; AsBo report; Safety Case	Certificates issued; Safety Case approved
8 – Authorization	Authorization decision; entry-into-service plan	Authorization granted; monitoring plan active
9 – Change and re-assessment	Change analysis; updated ODD and evidence	Change approved; residual risk acceptable

5.4.13 Change and update governance

During operation, updates may be required for the perception system, HD map or simulation models. Governance principles include:

- **Significance assessment:** Determine whether a change is significant (i.e., it affects safety or performance) under CSM RA and TSI rules. Significant changes trigger reassessment and may require updated authorization.
- **Stakeholder coordination:** Changes to infrastructure data (HD map updates) are initiated by the IM; changes to perception algorithms or models are initiated by the Supplier; changes to operational procedures or ODD are initiated by the RU. All changes require cross-stakeholder consultation.
- **Re-assessment triggers:** Examples include infrastructure modifications, detection of new hazard scenarios, model drift affecting detection performance, or changes in environmental conditions (e.g., new lighting or weather patterns). Monitoring data feeds into decision-making.
- **Documentation and communication:** Change logs, updated hazard logs, revised ODD specification and new verification evidence must be documented. Notifications to NoBo, DeBo, AsBo and ERA/NSA are required depending on change significance.

5.4.14 Exported constraints

The above-described responsibilities are just a first proposal for future work in the flagship area 2 or in the system pillar. In Wp6.4 not sufficient stakeholders could express their opinions about the acceptability of those constraints.

6 PROPOSAL FOR FURTHER STANDARDIZATION

6.1 Guiding principles

The guiding principles of further standardization are the following:

- Reuse first, horizontal standardization with other sectoral domain shall be prioritized unless rail-specific aspects have to be addressed (e.g. interfaces with rolling stocks, railway infrastructure,...)
- Prefer process + evidence patterns consistent with D6.4's Stage 0–9 workflow and EN 50716 constraints on ML in dependable software (compensating architecture and process measures).
- Track EU AI Act[20] standardization progress and avoid creating rail-specific requirements where an upcoming horizontal EN/ISO/IEC will suffice.

6.2 PRIORITY ACTIVITIES PRIOR TO FUTURE STANDARDIZATION ACTIVITIES

- Initiation of a standardization activity for dependable AI in the railway domain that extends the existing EN 50716 and that defines under which conditions AI may be used for dependable systems.
- PER systems using Ai for ATO GoA3/4 have to be considered as part of the category "high-risk AI systems" according to the EU AI act. Further work in JTC21 should therefore support the definition of a conformity assessment process to the EU AI Act applicable to PER System. [20]
- Define a standard for the "Certification Process" for perception systems for GoA3/4 systems in open environment based on the proposal from chapter 6. This might be prepared by the next activities in the innovation pillar of flagship area 2 from ERJU or may be transferred to the system pillar.
- Define a modular ODD for perception systems. This might be prepared by the next activities in the innovation pillar of flagship area 2 from ERJU or on a broader scope also for visual inspection systems via CEN/CLC TC 9X SG 38 AI.
- Recommend acceptance criteria for artefacts required to build perception systems for aspects like. This might be prepared by the next activities in the innovation pillar of flagship area 2 from ERJU.
 - Data quality of training, test and validation data sets including labeling guidelines according to the modular ODD definition
 - Required coverage of the dimensions of the ODD for validation providing sufficient evidence for dependable systems like PER.
 - HD-map quality & change management - Requirements for rail digital maps used by PER (content, accuracy, versioning, delta-updates), with safety-case hooks and operational change control.
 - Sufficient fidelity of the simulation & correlation with the real world that may be used for training, test and validation of the PER system. – This might be

- achieved by tailored fidelity classes (infrastructure, environment, sensors), accepted correlation methods of the simulations to field data, and acceptance of XiL/virtual testing as evidence
- Operational monitoring - Runtime monitoring requirements (out of operational design domain detection, drift, performance), logging (forensics & retraining), and controlled reassessment triggers

6.3 EXPECTED BENEFITS

If the proposed activities can be conducted and an agreement can be reached, the risk for the development of AI-based PER systems will be significantly reduced, because:

- The stakeholders can build up the required knowledge and skills and can reuse partially artefacts in a sequence of projects
- A tool and service ecosystem may grow that reduces costs supports reuse and risk reduction
- Increased transparency of the overall required artefacts for a PER based project may reduce the risk of under – or overestimations of the efforts to build and introduce a PER system in the market

7 IMPROVEMENTS & LIMITATIONS OF DELIVERABLE

The deliverable D6.4 has been established under the ground that functions based on perception contain AI algorithms. Yet, it can also be considered that safety-related functions of the Perception system would not depend on AI algorithms. This means of simplification of the certification process need also to be considered for Perception systems.

Regarding upcoming EU's Rail improvements, first, D6.4 deliverable still needs to be improved by a careful review of the System Pillar experts. This review would allow alignment of the proposed AI-based certification process with current regulations and standards.

The process should be mapped or aligned with the process workflow from EN 50126 such that it is easier to integrate it into the standard workflow of the system development process.

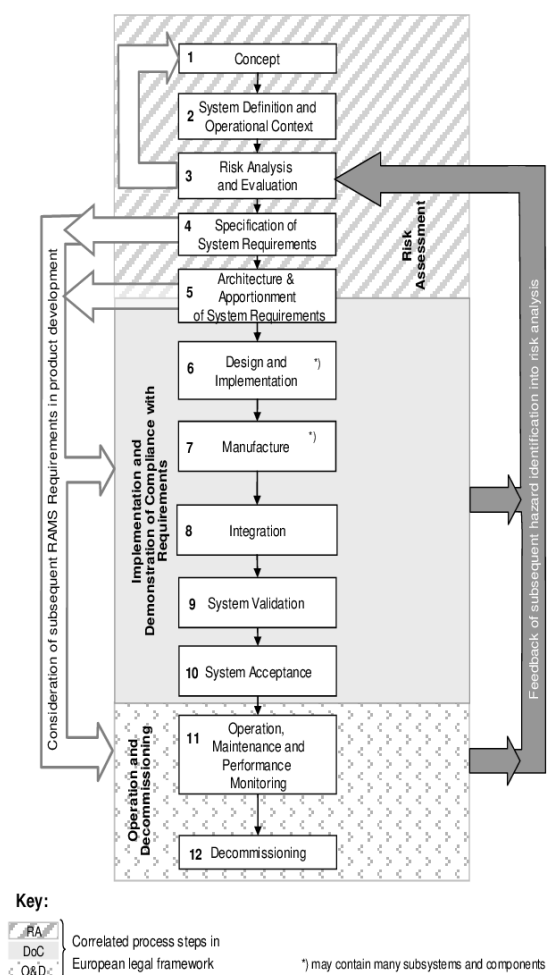


Figure 18: Interrelation of RAMS management process and system life cycle“ as defined by EN 50126 -12017

Regarding the standardization effort, relevant European and international AI standards have been identified but are not yet finalized. The TAURO D1.2 recommendation about horizontal standardization needs to be further pursued as accompanying standards of AI Act[20] are still under construction. Contribution of the railway sector in particular to ISO/IEC AWI TS 22440 on Functional safety and AI systems seems of great interest. Attention must be paid to EN50716 whose conclusion limit Artificial Intelligence and Machine Learning only to Based Integrity. Evidence must be provided, that training data are sufficiently complete and accurate, the trained software has been verified, approximated functions are validated, protection is granted against adversarial attacks and missing causality.

Regarding on-going initiatives, the state-of-the-art (TAURO[16], DEEL[1], Confiance.ai[15], Safe.TrAIIn[11]) projects, led to the identification of the major proposal and pitfalls of the certification of AI-based Perception system. Those elements have been noted by D6.4 to foster a sectoral contribution in on-going initiatives like DEEL or Confiance.ai. The main challenges of AI-based Perception system certification are data quality, robustness, explainability, uncertainty quantification and continuous monitoring. Common effort with those initiatives can foster successful outcomes on existing challenges.

Regarding the proposed process for Certification, the split of responsibilities between IM, RU and Suppliers did not lead to a consensus as some exported constraints on Users seem too complex to be achieved in a timely manner. The proposal of a PER-specific contract between RU, IM and Suppliers is a new approach for railway on-board equipment development which may also appear difficult to set up.

8 CONCLUSIONS

This deliverable proposes a certification process for optional AI-based perception (PER) systems on rolling stock that is consistent with railway regulation and aligned to the obligations for high-risk- AI according to the EU AI Act. It consolidates a staged workflow (Stages 0–9), defines roles and interfaces among ERA, NSAs, NoBos/DeBos, and AsBos, and links the process to the CSM for Risk Evaluation and Assessment. The approach is grounded in the architecture and requirements from WP 6.3 and the safety analysis from WP 8.2, making the dependencies to system specification and hazard results explicit.

Main contributions

1. **Process architecture.** Section 5 provides a general workflow and an extended, stagewise realization (Stages 0–9) from scoping and ODD definition through simulation/HiL/SiL, field validation, conformity and safety assessment, authorization, and change management. RACI tables and an evidence plan clarify responsibilities and deliverables across actors, reducing ambiguity in who creates, reviews, and accepts each artefact.
2. **PER-specific assurance.** Section 5.3 identifies PER critical artefacts and acceptance conditions: modular ODD, performance and false alarm metrics tied to safety goals, dataset and labeling governance, HD-map quality and change control, simulation fidelity, and an explicit safety case evidence map. These items operationalize how “PER-specific” evidence complements classical RAMS.
3. **Integrating requirements from law and standards into the certification process.** The process integrates PER and AI related aspects to rail use cases, like the EU AI ACT, ISO/IEC TR 5469 guidance on AI safety (data–learning–inference) and selects terminology/taxonomies from Inconsistent international standards for consistent rail-specific terminology/taxonomies by referencing to DIN DKE SPEC 99002/99004 (e.g., AI terminology or ODD).
4. **Evidence linkage to state of the art.** Using lessons from TAURO D1.2 (scenario-based V/W lifecycles, layered safety cases) and safe.trAI (PFD based acceptance tied to recall, nonconventional redundancy, SafeMLOps), the document explains how statistical performance, monitoring, and fallback measures can substantiate claims within the safety case. For certain analyzed functions, WP 8 indicates integrity needs up to SIL 2, contingent on architecture and barriers.

What the evidence shows

The combined “V+W” lifecycle with learning assurance, scenario libraries, and multimodal testing (virtual, SiL, HiL, field) is necessary to scale evidence for PER functions. Acceptance metrics must be traceable to safety objectives (e.g., linking recall to Probability of Failure on Demand), with uncertainty reporting and latency/availability treated as first class requirements. A modular, reusable ODD encoded against a common taxonomy increases repeatability and reduces project specific reinvention. These conclusions are reflected in the structure of Section 6 and its emphasis on artefact governance, milestone/evidence planning, and change control.

Constraints and open issues

Despite the proposed pathway, several constraints remain. First, normative gaps persist between AI guidance and rail standards; EN 50716 still cautions against AI for SIL1–SIL4 functions, which necessitates compensating architectural means (diversity, supervision, safe subsets) and careful partitioning of safety responsibilities. Second, regulatory timing is a factor: implementing measures and harmonized standards to operationalize the EU AI Act are evolving, which can affect conformity routes for high-risk AI used in rail. Third, demand mode interpretation for certain PER roles (continuous vs. low-demand emergency “parachute” functions) requires case-by-case alignment

with hazard analyses and external barriers. Finally, data and simulation fidelity (representativeness, correlation to field, and coverage of rare events) remain limiting factors for statistical confidence and must be managed through governance and continuous monitoring.

Implications for deployment

A pragmatic introduction path is to deploy perception initially with assistance level safety responsibilities, backed by dissimilar sensing/detection chains, runtime monitors, and explicit ODD supervision. Where WP 8 analyses indicate up to SIL 2 needs for specific PER linked safety functions, acceptance must be argued via quantitatively justified metrics (e.g., PFD linked to recall) and reinforced by architectural safeguards and independent assessment. The extended workflow's Stage 9 enables controlled updates (models, datasets, HD-maps) with reassessment triggers and versioned evidence to keep authorization valid over time.

Directions for the System Pillar and standardization

To reduce certification friction and support market adoption, the work suggests: (i) standardizing a machine-readable ODD schema and exchange format grounded in DIN KE SPEC 99004; (ii) defining dataset description and governance templates (provenance, coverage, labeling quality) traceable to acceptance metrics; (iii) harmonizing acceptance metrics and their mapping to safety goals for typical use cases; (iv) agreeing on scenario and evidence exchange interfaces for cross modal testing; and (v) formalizing role/RACI patterns for PER within conformity and safety assessment. These topics should be advanced through CEN/CENELEC TC 9X SG 38 (AI) with liaison to horizontal AI safety work (e.g., ISO/IEC TR 5469 and its successor TS series) and alignment to ERA processes and TSI evolution.

Conclusion

D6.4 provides a blueprint that integrates classical RAMS with AI specific assurance to certify perception functions in European rail. It defines how stakeholders can plan, generate, and maintain the evidence needed from ODD definition to authorization and through lifecycle change. The next project period should use this framework in pilots and System Pillar reviews to harden the artefacts, converge on acceptance metrics, and mature the standardization items listed above.

REFERENCES

- [1] The DEEL Github containing all developed libraries available at: <https://github.com/deel-ai>
- [2] The confiance.ai program, which notably aims to industrial DEEL libraries available at: <https://www.confiance.ai/>
- [3] FRSign, a French railway signalling dataset available at: <https://frsign.irt-systemx.fr/>
- [4] **European Union Agency for Railways (ERA)**, *Guide for the application of the locomotive and passenger rolling stock Technical Specification for Interoperability (LOC&PAS TSI)*, 2020. [European Union Agency for Railways](#).
- [5] Commission Implementing Regulation (EU) No 1302/2014, “Commission Implementing Regulation (EU) No 1302/2014 of 18 November 2014 on the technical specification for interoperability relating to the ‘rolling stock — locomotives and passenger rolling stock’ subsystem of the rail system in the European Union,” Official Journal of the European Union, L 356/228, 2014. [European Union Agency for Railways](#),
- [6] Commission Implementing Regulation (EU) No 402/2013, “Commission Implementing Regulation (EU) No 402/2013 on the common safety method for risk evaluation and assessment,” Official Journal of the European Union, L 121/8, 2013. [EUR-Lex \(EC\) No 352/2009](#)
- [7] IEC 61508-1:2010 Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 1: **General requirements**. [IEC 61508-1:2010](#)
- [8] ISO/IEC TR 5469:2024, Artificial intelligence — Functional safety and AI systems, first edition, Jan. 2024. [ISO/IEC TR 5469:2024](#)
- [9] [DIN DKE SPEC 99002:2025-03 - Terminology - AI in railway applications; Text in English](#), 2025
- [10] [DIN DKE SPEC 99004:2025-05 - Specification of Operational Design Domain in Rail; Text in English](#), 2025
- [11] Safe.trAIIn Projekt [safe.trAIIn](#)
- [12] safe.trAIIn Project, Project Outline – External Version, Jan. 2025. [ATRASS Seminar #5 safe.trAIIn - Safe AI for driverless regional trains](#)
- [13] [safe.trAIIn Report - Final Report \(German\)](#);
- [14] safe.trAIIn, [VEHITS25: Keynote Presentation- safe.trAIIn: Safety Assurance of a Driverless Regional Train](#)
- [15] Confiance.ai Consortium, Confiance.ai White Paper — Engineering Trustworthy AI, 2022. The programme aims to secure, certify and improve the reliability of AI systems; [Confiance.ai](#)

- [16] [TAURO Consortium, TAURO D1.2: Proposal for certification methodology of safe functions based on artificial sense, final version, 2024. CORDIS;](#)
- [17] ERJU R2DATO D6.3 – Architecture and System Specification for Perception System
- [18] ERJU R2DATO D8.2- Safety analysis for perception functions
- [19] ERJU R2DATO D27.2 – Specification of the Digital Register implementation [D27.1—Set-of-requirements-on-the-Digital-Register-in-R2DATO](#)
- [20] European Commission, “Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act),” COM/2021/206 final, Apr. 2021 [EUR-Lex -COM/2021/206 final](#)