Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26443-cd

Assurance Framework for Safe and Trustworthy AI in Railway Systems

Manuel Müller; Stefan Brunner; Leticia Fernandez Moguel; Monika Reif
Safety-Critical Systems Research Lab, Zurich University of Applied Sciences ZHAW, Switzerland.
E-mail: {manuel.mueller; stefan.brunner; leticia.fernandezmoguel; monika.reif} @zhaw.ch

Jiwon Shin
Stadler Signalling, Switzerland. E-mail: jiwon.shin@stadlerrail.com

Safety assurance for railway software is traditionally based on defined objectives supported by structured processes and documentation. These processes provide confidence in the correctness and reliability of deterministic, rule-based applications. With the introduction of artificial-intelligence-based perception and automation functions, this approach must be extended. The behavior of data-driven models cannot be fully demonstrated through conventional verification alone, and additional documentation, processes, and technical evidence are required to capture their development, validation, and operational control. Concurrently, new regulatory frameworks such as the EU Artificial Intelligence Act and emerging standards from CEN-CENELEC JTC 21 and ISO/IEC JTC 1/SC 42 introduce explicit requirements for fairness, robustness, transparency, traceability, and human oversight in high-risk AI systems. These provisions complement existing railway safety obligations but demand new methodological means of demonstrating compliance. To address these combined needs, this paper introduces a conceptual assurance framework for AI-enabled automated railway systems that extends process- and documentation-based assurance with technical and methodological measures. These are organized into coordinated pipelines covering the life-cycle of AI-enabled components: a data pipeline ensuring dataset quality and governance; a training pipeline managing model configuration and reproducibility; a verification and validation workflow combining scenario-based testing with complementary methods such as stress and sensitivity analysis, adversarial attacks, and uncertainty quantification to evaluate model behavior and robustness; and a monitoring pipeline enabling continuous assessment of performance and degradation in operation. Together, these pipelines address essential AI properties such as fairness, robustness, and transparency, translating regulation into traceable, auditable activities. The framework provides a consistent basis for aligning AI-specific compliance requirements with established railway assurance processes and generating evidence to support structured, transparent safety argumentation for future railway systems.

Keywords: Railway safety, AI assurance, Verification and validation, Regulatory compliance, Trustworthy AI

1. Introduction

Artificial intelligence (AI) and machine-learning (ML) technologies are increasingly being integrated into railway systems to enhance perception, decision-making, and automation capabilities. Typical applications include object detection and scene understanding in Communication-Based Train Control and driver-assistance systems, predictive maintenance and anomaly detection for rolling stock and infrastructure, and traffic-flow or energy-efficiency optimization (Xu et al. (2025); Chang et al. (2025); Hadj-Mabrouk (2024); Liu et al. (2024)). These technologies enable functions that can improve safety, availability, and operational efficiency (Liu et al. (2024); Oh et al. (2022)). Research and pilot implementations have demonstrated measurable gains in reliability and fault detection, yet the transition from experimental deployment to certified operation remains challenging (Labayen et al. (2023); Zeller et al. (2024)).

While these technologies promise substantial improvements, they also introduce new categories of risk. Among them are data drift and domain shift, which degrade model performance as operating conditions evolve. Moreover, the opacity of decision-making complicates traceability and verification. Further concerns include adversarial sensitivity, where minor perturbations lead to unsafe predictions, human-AI interaction issues, such as over-trust or inadequate situational awareness, and reward misalignment, where op-

timization objectives conflict with safety goals. Furthermore, systemic dependencies in which AI behaviors correlate across subsystems can amplify hazards (Schnitzer et al. (2024)).

Mitigating these risks requires assurance methods that go beyond traditional testing to include robustness evaluation, sensitivity analysis, explainability studies, and continuous operational monitoring (Labayen et al. (2023); Zeller et al. (2024); Roßbach et al. (2024)). Conventional railway software assurance, as defined by the *EN 50716:2023* standard, specifies structured processes for software development, verification, validation, and life-cycle management for rule-based, deterministic systems. However, AI components are data-dependent and probabilistic, as their performance depends on dataset quality, model configuration, and environmental variability. Such uncertainty cannot be fully addressed through traditional verification and validation techniques. For this reason, new regulatory and standardization initiatives are reshaping assurance expectations. The EU Artificial Intelligence Act (EU AI Act) (European Commission (2024)) classifies railway-related AI as high-risk and mandates requirements for transparency, traceability, robustness, human oversight, and post-market monitoring. Complementary international standards such as *ISO/IEC TR 24028:2020* define conceptual models of AI trustworthiness across the life-cycle. Together, these frameworks call for assurance approaches that integrate structured process control with AI-specific technical evidence. Recent research has made tangible progress toward this goal. Zeller et al. (Zeller et al. (2024); Zeller and Schnitzer (2025)) propose life-cycle-oriented MLOps processes for the continuous assurance of ML components, while Labayen et al. (Labayen et al. (2023)) and Roßbach et al. (Roßbach et al. (2024)) demonstrate hybrid V&V techniques combining simulation-based and data-driven testing. In parallel, cross-domain initiatives introduce trustworthiness frameworks that combine technical, organizational, and procedural evidence to support certification in safety-critical applications (Billeter et al. (2024); Frischknecht-Gruber et al. (2025)).

Despite these advances, a consistent methodology linking AI trustworthiness principles with the established, process-oriented assurance structure of the railway domain remains underdeveloped. This paper addresses this gap by introducing a conceptual assurance framework for AI-enabled railway systems. The framework translates AI assurance principles into railway-specific processes, aligning data governance, model training, scenario-based verification, and operational monitoring with the life-cycle defined in *EN 50716:2023* and the compliance expectations of the *EU AI Act*.

2. Regulatory Background

Railway safety assurance is grounded in structured, lifecycle-oriented processes defined by the *CENELEC railway standards series*. *EN 50126* establishes the RAMS life-cycle framework, while *EN 50129* defines the evidential structure of the safety case and the independence of assessments. The more recent *EN 50716:2023* complements these foundations by updating the software-development requirements for control and protection systems to reflect current engineering practice and to align with contemporary assurance methodologies (Tonk et al. (2023)). However, AI constituents challenge the assumption of deterministic behaviour and therefore require additional assurance measures that capture data quality, model robustness, and performance drift over time (Labayen et al. (2023); Xu et al. (2025)). Current standardization work within *CEN-CENELEC JTC 21* and the recently published *ISO/IEC PAS 8800:2024* explores how these AI-specific forms of evidence can be combined with established functional-safety life-cycles.

The regulatory landscape for AI Assurance is evolving rapidly. The *EU AI Act* introduces a risk-based regulatory framework that mandates documentation, testing, human oversight, and post-market monitoring for high-risk AI systems such as those used in railway automation. In parallel, international standardization bodies have developed complementary initiatives: *ISO/IEC TR 24028:2020* provides an overview of AI trustworthiness; *ISO/IEC 42001:2023* de-

fines requirements for AI management systems; *ISO/IEC 5338* and *5339* specify life-cycle and engineering practices for trustworthy AI; and *ISO/IEC PAS 8800:2024* addresses the integration of AI with functional-safety processes. Work in *CEN–CENELEC JTC 21* seeks to harmonize these frameworks with existing railway RAMS and safety-case methodologies.

Together, these standards expand assurance beyond conventional software verification to encompass the governance of data, robustness of models, transparency of decisions, and continuous oversight during operation (Billeter et al. (2024); Frischknecht-Gruber et al. (2025)). Yet their incorporation into railway certification practice remains at an early stage. There is still **no established mechanism** for systematically linking AI-specific artifacts, such as dataset lineage records, validation metrics, or uncertainty analyses, to the structured evidence required under CENELEC safety cases.

As a result, existing regulatory and standardization frameworks define what must be assured for AI systems in railways, but provide limited guidance on how such assurance can be operationalized within established railway certification practices.

Research gap and opportunities. Despite steady progress, assurance practice for AI-enabled railway systems remains fragmented across technical, organizational, and regulatory dimensions. Classical safety standards offer strong process discipline but presuppose deterministic behavior, while AI assurance methods emphasize data governance, robustness, and explainability in ways that rarely align with existing certification logic. The result is a structural mismatch: document-centric and requirement-driven assurance approaches versus empirical, statistical, and often non-repeatable forms of AI validation evidence. Bringing these perspectives together, so that data quality metrics, model-validation results, and uncertainty analyses can support structured safety arguments, remains an unresolved task.

Recent contributions by Zeller et al. (Zeller et al. (2024); Zeller and Schnitzer (2025)) demonstrate continuous-assurance and MLOps processes for ML components, providing valuable technical mechanisms and treating regulatory integration separately. Cross-domain trustworthiness frameworks (Billeter et al. (2024); Frischknecht-Gruber et al. (2025)) offer comprehensive assurance models for AI but stop short of linking them to sector-specific certification schemes. The methodological gap therefore lies between functional-safety-oriented standards such as *EN 50716* and *ISO/IEC PAS 8800*, and AI-trustworthiness frameworks such as *ISO/IEC 42001, 24028, 5338, and 5339*.

What is missing is an integrated approach that connects these two domains in a traceable and auditable manner. Such a framework must (1) translate AI trustworthiness concepts into railway life-cycle activities; (2) align data, model, and operational artefacts with established documentation and verification requirements; and (3) support continuous, auditable assurance consistent with both *EN 50716* and the *EU AI Act*.

This paper addresses that need by extending established railway-assurance practice with AI-specific verification, validation, and monitoring pipelines, providing a foundation for certifiable and trustworthy AI in future railway systems.

3. The Assurance Framework for Safe Railway AI Systems

The proposed framework builds upon established system engineering practices, particularly the V-model approach standardized in *SN EN 50126-1* for railway applications, while incorporating AI-specific methodologies from domains such as aviation, medical technology, and autonomous systems development. The framework supports five phases: (1) the concept phase, (2) the risk assessment and requirements definition, (3) the design and implementation phase, (4) the validation phase, and (5) the continuous monitoring and improvement. Fig. 2 visualizes the suggested framework.

The life-cycle starts with the **concept phase** defining intent and context. The *AI concept* parallels the traditional one but requires explicit consideration of AI-specific factors including the machine learning class and autonomy level. The sub-

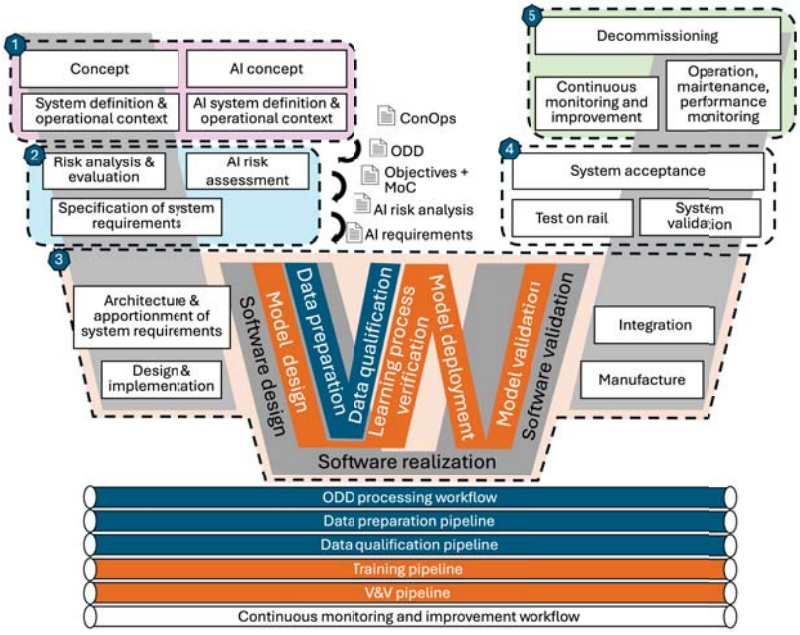

Fig. 1. Overview of the proposed assurance framework. The railway V-model lifecycle is extended with an AI constituent lifecycle, including ODD processing, data preparation and qualification, model training and validation, and continuous monitoring and improvement, implemented through coordinated pipelines and workflows.

sequent *AI system definition & operational context* focuses on the heterogeneous environments in which AI systems operate. While traditional systems engineering focuses on use cases representing functional requirements, AI systems require scenario-based analysis that captures the combinatorial complexity of environmental conditions. Unlike traditional systems where operational contexts are highly controlled, AI systems must function across large parameter spaces with semi-controllable environmental conditions. The *Concept of Operations (ConOps)* document captures these specifications in a traceable manner, whereas the *Operational Design Domain (ODD)* encompasses environmental elements (persons, vehicles, obstacles), parameters (weather conditions, velocities, poses), and their combinations into scenarios. It defines the boundary conditions for system behavior and guides training data collection strategies.

The subsequent **risk assessment and requirements definition** formalizes the needs identified in the concept, derives risks, and provides mitigation strategies. The *AI risk assessment* extends traditional risk analysis by explicitly addressing trustworthiness alongside safety concerns, including bias, lack of robustness, poor generalization, and potential for reward hacking. The framework recognizes that AI systems introduce such unique failure modes that stem from their data-driven and probabilistic nature, and documents it in the *AI risk analysis*. To manage the complexity of multiple, sometimes overlapping regulatory demands, the framework introduces the concept of Means of Compliance (MoC) alongside objectives as a central organizing principle, inspired by aviation safety assurance practices. MoCs serve as a bridge between abstract regulatory objectives and concrete implementation activities. Objectives capture the core intent of the regulations while the MoCs specify methods and recipes for demonstrating compliance. This approach enables systematic handling of the fragmented regulatory landscape including domain-

specific standards (e.g., *EN 50716:2023* for rail), AI-specific regulations (EU AI Act), and cross-domain frameworks (ISO/IEC standards series). In cases where different legal frameworks address similar concerns using different terminology or different procedures, the objectives of the legal frameworks are clustered and assigned to a few MoCs to ensure compliance with the entire regulatory landscape. As MoCs for AI constituent in safety-critical applications are momentarily scars compared with aviation industries' mature libraries, the framework provides guidance for developing new MoC subject to competent authority approval. The *system requirement specification* transforms the results of the risk assessment and the specified characteristics into concrete and verifiable requirements while considering the multiple regulatory sources applicable to the specific application domain, and documents it in the *AI requirements*.

The **design and implementation phase** complements the classic V-model's software design, realization, and validation (Fig. 1, gray) with the AI-specific model life-cycle (orange) and the data life-cycle (blue). This takes into account the parallel consideration of data, model, and software workflows, which are discussed in more detail in Section 4. The outcome of the design and implementation phase is an integrated and deployed AI constituent, which is rigorously assessed in the *Validation phase* .

The **validation phase** starts with *Test on rail*, where pilot programs assess the object and event detection and response capabilities in the real system. Drawing inspiration from clinical trials that manage high uncertainty similar to AI deployments, they evaluate ODD coverage and judge context-dependent benefit-risk balance. A typical pilot project was running the AI constituent in shadow mode operation parallel to operational systems without affecting live operations. This enables safe validation of AI behavior while human operators maintain control. Such extensive pilot test phases generate evidence supporting null hypothesis testing about system safety and performance claims. A report documents the validation results including hypothesis test outcomes and confidence levels achieved. This evidence forms the basis for the *system acceptance* step and provides input for operational monitoring strategies.

Continuous monitoring and improvement concludes the life-cycle. It runs in parallel to the operational phase and implements accident and incident tracking, data monitoring and analysis, V&V activities, user support, and explanation capabilities. Stemming from the regulatory landscape, several AI-specific monitoring activities complement the traditional ones. An ethics review board addresses value alignment and fairness concerns. The input stream is compared against predefined regions where generalization capability maintains intended behavior. Data-driven safety checks complement traditional rule-based monitoring. This monitoring infrastructure feeds continuous improvement cycles, enabling data-driven refinement of models and updating of operational parameters based on field experience.

To automate the activities in the life-cycle, four pipelines and two workflows are associated with the framework. Both, the *ODD processing workflow* and the *Continuous monitoring and improvement workflow* involve user interaction to create the required simulation and data capturing artifacts and to provide issue reports, as well as accepting suggested improvements. In contrast, the *Data preparation pipeline*, the *Data qualification pipeline*, the *Training pipeline*, and the V&V pipeline run with a very high degree of automation.

4. AI Design and Realization Workflow

As software design, realization, and validation for AI resemble the classic software development process, this section focuses on the data and model workflow, as visualized in Fig. 2. The framework explicitly separates, yet coordinates, data engineering and model training as two distinct development streams.

This reflects the fact that they have different quality criteria and verification needs, while also acknowledging their fundamental interdependence. Consequently, the *learning process planning*, the *model training*, and the *learning process verification* serve as an integration point, whereas

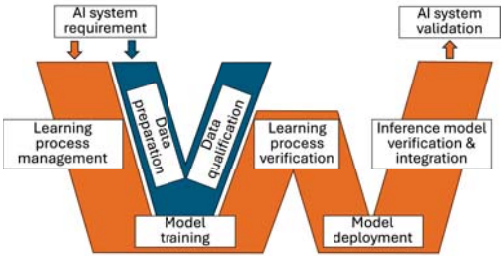

Fig. 2. Workflow for data and model development, coordinating data engineering and training to produce traceable assurance artifacts.

the other steps are distinct.

The workflow starts with the *Learning process management*, which formalizes goals to ensure the AI constituent aligns with intended objectives and stakeholder needs. It contains the data strategy and the learning strategy. The data strategy specifies comprehensive validation criteria including ingestion protocols, perturbation methods, and statistical properties such as dataset completeness, accuracy, and representativeness. The learning strategy incorporates confidence level estimation and systematic risk mapping. Linking key performance indicators, decision boundaries, and cybersecurity vulnerabilities to potential failure modes supports risk mitigation for robust model development and deployment. Learning process management represents a departure from traditional software design, reflecting the experimental nature of AI development. Rather than specifying system behavior through detailed design, this phase involves conducting a systematic literature review to identify potential architectures. The most promising candidates are selected for experimental validation by clustering and comparing them using methods such as SWOT analysis.

The *Data acquisition* runs alongside. It encompasses ingestion, integration, and data augmentation. The process begins with planning concepts, modalities, and scenarios to be captured, linking them directly to the ODD specifications and ensuring that training data coverage matches the intended areas of application. To this end, synthetic data that contains rare environmental effects will complement the data captured from real-world applications. Data augmentation increases the size and coverage of the effective dataset, and active variant generation explores robustness boundaries. This approach is similar to mutation testing, but addresses the statistical rather than the deterministic nature of AI systems. The framework recognizes simulation as a critical tool for AI systems, enabling systematic exploration of scenario spaces that would be impractical or unsafe to capture in operational environments. However, simulation-to-reality transfer introduces validation challenges that must be addressed through careful characterization of domain gaps.

Having selected the baseline architecture and collected and augmented the data, the *AI training* optimizes the model weights to align the behavior with the intended functionality. This iterative process gradually improves the model's capability to handle increasingly complex situations within the defined ODD.

The subsequent *Learning process verification* employs multiple verification strategies: bias-variance analysis to detect over- or under-fitting, performance metrics (correctness, accuracy, robustness, generalization, stability, etc.), generalization bound estimation, and requirements-based behavior checks. These evaluations serve similar purposes to software verification but require interpretation through the lens of statistical confidence rather than deterministic correctness. The results are documented in the model portrait, which provides a consolidated view of model characteristics including architecture, learning algorithms, explainability capabilities, activation and loss functions, robustness and stability metrics, initialization strategy, training environment, and hyperparameters. Analogous to a hardware datasheet, the model portrait enables informed decisions about model suitability. The *Data qualification* assesses completeness, accuracy, representativeness, and statistical properties against predefined criteria from the learning process plan. This verification step addresses a fundamental challenge in AI development: anticipating which features of the dataset will prove decisive for model behavior, given that AI perception differs substantially from human perception. The framework specifies crite-

ria including sufficient sample count, appropriate statistical balance, correct formatting for training pipelines, and proper partitioning into training, test, and validation sets. Crucially, these sets must maintain strict separation to enable valid performance estimation. The resulting *Data Quality Report* documents compliance with these criteria. The *learning process verification* merges data and model V&V workflows, ensuring that both data quality goals and model performance goals have been met before progression to deployment. The process checks the achieved results against expectations from learning process planning, and reviews completeness.

Once the learning process concludes successfully, the *AI deployment* transforms trained models into operational systems through API development, format conversion wrappers, and containerization. The framework suggests release candidates as optimization-specific packages including inference-optimized models without training-specific operations like gradient calculation, wrappers between AI standard formats and domain-specific formats, and potentially modified execution infrastructure. These candidates undergo *Inference model verification & integration* in sandboxed environments before being tested on rail.

5. Discussion and Outlook

The Assurance Framework for Safe Railway AI Systems addresses critical challenges in developing trustworthy AI for safety-critical applications by translating them into railway life-cycle activities. The Means of Compliance concept enables systematic demonstration of regulatory compliance across fragmented standards by identifying commonalities, developing unified approaches, and aligning them with established documentation and verification requirements. The explicit separation and coordination of data, model, and software engineering takes into account the characteristics of AI. The data pipeline demonstrates systematic coverage in combinatorially large operational spaces by linking ODD parameters to scenario generation, data collection, and validation. By emphasizing statistical approaches and confidence estimation, the V&V workflow acknowledges that AI systems cannot achieve deterministic correctness and instead building evidence-based confidence while explicitly recognizing remaining uncertainties. Documentation artifacts establish crucial traceability between requirements, implementation decisions, and validation evidence. In this way, the framework supports continuous, auditable assurance consistent with both railway-specific regulatory frameworks, and AI-specific ones.

However, several challenges remain unresolved. The scarcity of established means of compliance requires projects to develop and gain acceptance for them in a case-by-case manner, potentially creating inconsistent practices until consensus emerges. Computing meaningful generalization bounds in complex, high-dimensional spaces remains an open research challenge, forcing reliance on empirical validation whose adequacy may be difficult to demonstrate. Fundamental tensions between model performance and interpretability persist, potentially requiring reduced capability to maintain transparency in safety-critical contexts. The framework focuses primarily on fixed, validated models and would require additional safeguards for systems with continuous learning. Comprehensive verification activities demand substantial resources, necessitating risk-based tailoring for practical applications. The framework's grounding in established safety engineering practices while incorporating AI-specific needs offers organizations a practical path for system certification and deployment. Its elaboration of similarity and difference enable efficient knowledge transfer from traditional domains while preventing inappropriate application of classical methods. Future work should develop domain-specific means of compliance libraries, refine generalization bound estimation methods, extend the framework to accommodate continuous learning, develop supporting tooling, and conduct empirical validation across diverse applications.

Acknowledgement

This work was conducted within the Innosuisse-funded project VRAIL, a collaboration between Stadler Signalling and ZHAW.

References

Billeter, Y., P. Denzel, R. Chavarriaga, O. Forster, F.-P. Schilling, S. Brunner, C. Frischknecht-Gruber, J. Weng, and M. Reif (2024). Mlops as enabler of trustworthy ai. In *11th IEEE Swiss Conference on Data Science (SDS)*, pp. 37–40.

Chang, C.-C., K.-H. Huang, T.-K. Lau, C.-F. Huang, and C.-H. Wang (2025). Using deep learning model integration to build a smart railway traffic safety monitoring system. *Scientific Reports 15*(1), 4224.

European Commission (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence (AI Act). Official Journal of the European Union, L 202, 12 July 2024, pp. 1–105. Entered into force 1 August 2024; applicable from 2 August 2026.

Frischknecht-Gruber, C., P. Denzel, M. Reif, Y. Billeter, S. Brunner, O. Forster, F.-P. Schilling, J. Weng, and R. Chavarriaga (2025). Ai assessment in practice: implementing a certification scheme for ai trustworthiness. In *Symposium on Scaling AI Assessments (SAIA 2024)*, Volume 126 of *OpenAccess Series in Informatics (OASIcs)*, pp. 15:1–15:18.

Hadj-Mabrouk, H. (2024). A literature review on the applications of artificial intelligence to european rail transport safety. *IET Intelligent Transport Systems 18*(12), 2291–2324.

Labayen, M., X. Mendialdua, N. Aginako, and B. Sierra (2023). Semi-automatic validation and verification framework for cv&ai-enhanced railway signaling and landmark detector. *IEEE Transactions on Instrumentation and Measurement 72*, 1–13.

Liu, J., G. Liu, Y. Wang, and W. Zhang (2024). Artificial-intelligent-powered safety and efficiency improvement for controlling and scheduling in integrated railway systems. *High-speed Railway 2*(3), 172–179.

Liu, J., X. Sun, F. Wang, W. Zhang, and B. Yu (2024). Artificial-intelligent-powered safety and efficiency for composite railway systems. *Journal of Rail Transport Planning & Management 29*, 100573.

Oh, K., M. Yoo, N. Jin, J. Ko, J. Seo, H. Joo, and M. Ko (2022). A review of deep learning applications for railway safety. *Applied Sciences 12*(20), 10572.

Roßbach, J., O. De Candido, A. Hammam, and M. Leuschel (2024). Evaluating ai-based components in autonomous railway systems: A methodology. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pp. 190–203. Springer.

Schnitzer, R., L. Kilian, S. Roessner, K. Theodorou, and S. Zillner (2024). Landscape of ai safety concerns – a methodology to support safety assurance for ai-based autonomous systems. In *2024 8th International Conference on System Reliability and Safety (ICSRS)*, pp. 501–510. IEEE.

Tonk, A., S. Wimmer, A. Turetta, T. Groß, and M. Ulbrich (2023). A safety assurance methodology for autonomous trains. *Safety Science 172*, 105628.

Xu, X.-Y., S.-M. Wang, W.-Q. Liu, and Y.-Q. Ni (2025). Advancements in obstacle intrusion detection methods for rail transit: A comprehensive review. *IEEE Transactions on Instrumentation and Measurement*.

Zeller, M. and R. Schnitzer (2025). Assuring safety of ai-based systems: Lessons learned for a driverless regional train case study. In *Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference*, pp. 1411–1418. Research Publishing, Singapore.

Zeller, M., T. Waschulzik, R. Schmid, and C. Bahlmann (2024). Toward a safe mlops process for the continuous development and safety assurance of ml-based systems in the railway domain. *AI and Ethics 4*(1), 123–130.