Digital Society (2022) 1:17
https://doi.org/10.1007/s44206-022-00015-1

ORIGINAL PAPER

Auditing of AI in Railway Technology – a European Legal Approach

Dagmar Gesmann-Nuissl[1] · **Stephan Kunitz**[1]

Received: 7 March 2022 / Accepted: 28 July 2022 / Published online: 2 September 2022

Abstract
Artificial intelligence (AI) promises major gains in productivity, safety and convenience through automation. Despite the associated euphoria, care needs to be taken to ensure that no immature, unsafe products enter the market, especially in high-risk areas. Artificial intelligence systems are therefore to be integrated into the European Union's existing product safety system by the planned AI regulation. This is accomplished by horizontally linking the draft AI regulation (AI-Act) with the existing harmonizing legal acts that ensure the safety of products and provide for a complex system of approval and testing concepts for this purpose. The railway sector is no exception, which is why potential AI systems for monitoring tracks or simple and accurate train detection are also subject to this approval regime. The following article highlights the challenges that exist in the railway sector on the path to verifiable AI systems in this regulatory context.

Keywords Artificial intelligence · Auditing · Railway · Conformity assessment · Proof of safety

1 Introduction

Railway systems are expected to become safer and more reliable through AI-based applications. The integration of AI promises a more precise object and obstacle detection (OOD) in the track environment compared to the human gaze, as well as more accurate and faster determination of train positions and completeness of the train, e.g. by using fibre-optic sensing (FOS). Due to this expected precession, such AI-based or AI-assisted applications will become

✉ Dagmar Gesmann-Nuissl
dagmar.gesmann@wiwi.tu-chemnitz.de

✉ Stephan Kunitz
stephan.kunitz@wiwi.tu-chemnitz.de

[1] Faculty of Economics and Business Administration, Chemnitz University of Technology, Reichenhainerstr. 41, Chemnitz 09126, Germany

significantly more important in the digitization and automation of railway transport. The envisioned driverless and resource-efficient mobility of the future and the integration of railway transport into smart mobility concepts will depend on the rapid integration of these AI-based applications.

However, this expectation does not coincide with the current requirements of the European railway regulations for the functional safety of railway systems and their verification procedures. They are not prepared for AI-based applications. Therefore, the development of testing standards and methods for using AI in the context of innovative railway systems is required, which ultimately also have to be compatible with the draft European Artificial Intelligence Act (AI-Act).

A particular challenge in this context is the procedure for the approval-relevant proof of safety, especially if the decision of the AI is not or only insufficiently comprehensible and explainable due to its training data or if ethical aspects suddenly have to be included in automated decision processes. This proof of safety is one of the results of a conformity assessment conducted to demonstrate the required functionality and safety of a product. Since these assessments aim to prove safety and functionality, they serve as a kind of audit in a broader sense. Due to the need for conformity assessments and the proof of safety, this article will focus on these types of (technical) audits.

The extent to which strategies for deriving and implementing acceptance criteria for machine learning (ML) functions can support safety verification and how the certification and approval of the required systems could function will be shown after the fundamental challenges of the current regulatory environment.

In this paper, we attempt to present the current problems in the safety verification of AI systems in the railway sector. A major challenge is the lack of transparency inherent in the use of AI, which is commonly described as the "black box problem". The higher the sector-specific requirements for proving the functional safety of a (high-risk) system are, i.e. the more precisely these processes have to be described, the more difficult it is to provide this proof of safety. However, it remains mandatory to launch products on the market or on rail.

In an understandable way, we will describe the challenges that arise from the interaction between the AI law and sector-specific regulations. We also suggest some possible solutions to surmount this problem. In doing so, the paper is structured as follows: After a brief introduction to AI and its importance for automation in the railway sector, two pioneering AI applications are presented in detail in Sect. 2. Section 3 discusses the legal framework regarding product safety and regulation in the European Union. Here, we focus on describing how the concept of the recently proposed AI law and sector-specific regulations intertwine under the umbrella of the new legislative framework (NLF). In Sect. 4, we provide a deeper immersion into the sector-specific railway regulations, showing where and how the "black box problem" inherent in AI impacts the sector-specific regulations for demonstrating functional safety. Section 5 summarizes the results of the previous sections and suggests how the problem of the proof of safety for AI-related applications could be solved.

2 Artificial Intelligence in Railway Technology

2.1 Artificial Intelligence

The concept of artificial intelligence (AI) cannot be clearly delineated due to the numerous attempts at definition that have been made so far in the various scientific disciplines (Kaulartz & Braegelmann, 2020; Russel & Norvig, 2012; Wang, 2019). However, it can be stated that the term AI hides several specific technologies, such as machine learning concepts in the shape of neural networks, logic- and knowledge-based concepts and statistical approaches or Bayesian estimation, search and optimization methods. All these concepts have in common that they design and optimize systems, arrangements or decisions more or less autonomously on the basis of algorithmic commands and oriented on biological models. They operate in a self-learning, reasoning and self-correcting manner (Bittner et al., 2021). The AI-Act, which set out with the objective of regulating AI, includes these various technologies in its Annex I and is sufficiently receptive to further (future) concepts (Art. 4 AI-Act, Art. 290 Treaty on the Functioning of the European Union). In the following, the approach is based on a very broad understanding (Osborne, 2021) of the term AI, which is based on the definition of the AI Act. According to this definition, an AI system is "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (Art. 3 No. 1 AI-Act).

2.2 Grade of Automation and AI-Based Applications

According to the aforementioned definition, several applications and systems are currently being designed in the context of railway transport. On the one hand, AI systems are intended to automate repetitive tasks to enable rail-specific processes to be completed faster and better than by a human (e.g. AI-based OOD). On the other hand, they are intended to help increase the precision of traffic flow to enable more cost- and energy-efficient operations in the future (e.g. AI-based determination of train position by FOS). Both applications will be explained in more detail after a brief description of the automation levels in railway transport.

In railway transport – similar to other mobility sectors – there are different levels of automation (Grade of Automation, GoA), which are defined in more detail in the IEC 62267 Standard. The grades of automation start with GoA 0, conventional driving on sight, without any automatic train protection that could serve as a fallback. GoA 1 describes manual driving with automatic train protection. Train operation as such is not yet automated, but the driver performs all driving functions manually. GoA 2 describes semi-automated train operation involving a driver. At this level, the train control system is supplemented by automatic traction and automatic braking. Although the train runs automatically, the driver can take over control at any time and switch to manual driving. GoA 3 refers

to autonomous driving with on-board personnel. The system handles all driving functions fully automatically, including environment monitoring such as signal and obstacle detection. In this driverless train operation, there is no longer any control by an engine driver. The train is accompanied only by on-board personnel who are responsible for door control and emergency operation. Fully automatic, driverless train operation is achieved with GoA 4. In this stage, there is no longer any on-board personnel on the train; all processes run fully automated. It is only the control centre that can still intervene in train operation.

In rail operations according to GoA 2, where the common rail traffic is located today, the driver is required by a service regulation to observe the track while the train is in motion in order to detect hazards at an early stage. Hazards include people, animals or objects on or beside the track, as well as malfunctioning overhead lines and signalling equipment. However, in the future, at higher levels of rail automation (grade of automation, GoA 3 and 4), when increased automation of rail operations (automation levels, GoA 3 and 4) means that the train driver will no longer be on board, and track monitoring will have to be replaced by reliable technical systems. This type of automation of train operations (ATO) is currently one of the most important trends in railway development, not least to address the strong competition from other modes of transport (Mockel & Scherer, 2003; Reinhold & Kasperkovitz, 2013; Ristić-Durrant et al., 2021; Weichselbaum et al., 2013). While many components for automatic train operation have already been developed (e.g. automatic train control, automatic door control, and automatic departure from the station), driverless train operation (GoA 4) has been enabled only for systems in selected environments (e.g. subway systems and mining lines in uninhabited areas). One of the major problems in implementing driverless trains on public railway infrastructure is the risks associated with undetected objects and obstacles on the track.

2.2.1 Object and Obstacle Detection

Therefore, the implementation of a train-side camera system with an AI-based image recognition is discussed for OOD (Braband & Schäbe, 2020; Braband, 2021; He et al., 2021; Yu et al., 2018; Dagvasumberel et al., 2021). Their use is intended to provide security in a manner that only humans have been able to do. To this end, the cameras take images or sequences of images of the track and immediate surroundings at regular intervals. In order to detect and classify obstacles and determine their precise position on the track, the distance from the train or the size of the obstacle, an AI-System previously trained on obstacles is continuously applied to these images. This system evaluates the image data and provides its results, including existing uncertainties, to a local or remote system. This system finally initiates the appropriate reaction (braking the train, notifying the following train or emergency services; see Fig. 1), considering other information such as train positions and speeds. This data-based process operates in fractions of a second so that the result obtained could be equal to a human estimate. Since machine learning systems based on neural networks have proven to be remarkably effective in the field of image recognition, obstacle detection should

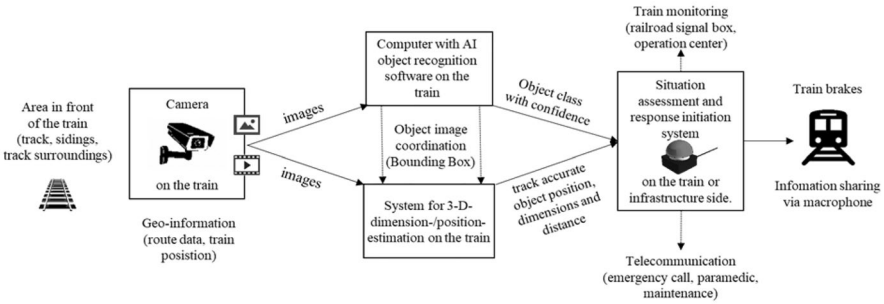

Fig. 1 AI-based OOD (adapted from Schwencke, D., Deutsches Zentrum für Luft und Raumfahrt, Project "KI-bezogene Test- und Zulassungsmethoden – Anwenderkompass für den intelligenten Schienenverkehr", SRCC)

be based on this type of technology. Accordingly, this would be a system that falls under the AI concept of Art. 3 (1) in conjunction with Annex 1 (a) AI-Act.

2.2.2 Fibre-Optic Sensing (FOS)

Another area of AI application is determining the position of trains for driving at block distance. Only once the train ahead has completely left a track section (with all waggons) and this is communicated to the track control system by a signal, the track can be cleared for the following train. The "moving block" concept can thus be defined as an automatic control system that allows each train to receive "movement authorization information" from the control centre in a safe mode. The control centre manages this process by being in constant dialogue with all trains and being informed of their speed and position. This safety procedure helps to avoid collisions between trains and ensures more efficient and thus more climate-friendly traffic management.

Technologically, the detection of train position and completeness is currently realized by recurring axle counting points in the track or by balises laid in the track. The maintenance of these technical facilities has proven to be extremely cost-intensive. Furthermore, these facilities do not generate any further added value. For this reason, research is being conducted on alternatives that also enable, for example, a telecommunication connection or maintenance monitoring (Kowarik et al., 2020; Vidovic & Maschnig, 2020; Vidovic & Landgraf, 2018). Fibre-optic sensing (FOS) is such an alternative. Fibre-optic sensors operate based on the principle that light is sent from a laser or other superluminescent source through an optical fibre laid along the track. Vibrations and shocks generated by a passing train cause variation in the signal parameters. These variations are measured in a detector and allow inferences to be drawn about the block length, speed and completeness of the train. The position, completeness and speed of the train can be determined very accurately using FOS (Kowarik et al., 2020).

AI could accelerate the generation and processing of sensor data in the context of FOS and thus determine or calculate the arrival and position of trains faster and more precisely than conventional peaking systems (Kowarik et al., 2020) (Fig. 2). Precise and rapid calculation of track utilization allows capacity on the track to be increased. In addition, incorrect – because colliding – position reports and thus train collisions can be avoided.

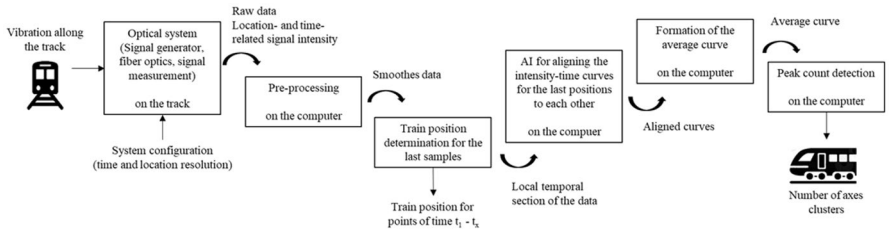

Fig. 2 AI-based FOS (adapted from Schwencke, D., Deutsches Zentrum für Luft und Raumfahrt, Project " KI-bezogene Test- und Zulassungsmethoden – Anwenderkompass für den intelligenten Schienenverkehr", SRCC)

3 Legal Framework

The aforementioned AI-based solutions have to fit into the security architecture of the European regulatory framework. Above all, they have to fulfil the product- or process-specific safety requirements in order to obtain approval for the European Single Market.

3.1 New Legislative Framework

The new legislative framework (NLF) provides the European Single Market with a system of product safety and surveillance based in particular on the participation of manufacturers. For this purpose, the EU Commission defines the basic safety requirements for products in European directives. The technical design of the product-specific safety requirements is subsequently the responsibility of private standardisation bodies. These are mandated by the EU Commission to develop the necessary technical standards that further specify the minimum requirements from the directives (Ensthaler & Gesmann-Nuissl, 2006; Ensthaler et al., 2012; Kapoor & Klindt, 2008). References to harmonized standards need to be published in the *Official Journal of the European Union*.

Products manufactured in accordance with the so-called European harmonized standards are considered to be in conformity with the directive. It is assumed that the products comply with the respective essential safety requirements (Art. 3 (2) 1 Directive 2001/95/EC). The product receives the CE marking as an external sign. In general, there are two ways of conducting these types of audits for compliance with the harmonized standards. Depending on the degree of risk of the respective product, compliance with the essential requirements is audited either by the manufacturer themselves or by a so-called notified body (NoBo) as an independent test facility with special competence. A so-called conformity assessment procedure is conducted (Ensthaler et al., 2012; Kapoor & Klindt, 2008). If the respective product has to fulfil special safety requirements – regularly those of functional safety – a separate proof of safety is also required as part of the conformity assessment. Similar to the proof of conformity, this proof of safety has to be provided by the manufacturer or an independent, certified body.

The European regulatory framework for the railway sector remains extremely heterogeneous due to interoperability. However, the basic rail regulations widely follow the NLF. In the case of railway-specific regulations too, a series of directives provide the essential guidelines, which are then supplemented by individual standards and passed on to the member states for national implementation. Which specific regulations apply to the definition of safety requirements or conformity assessment depends on which subsystem is to be approved under Annex II of Directive (EU) 2016/797. A distinction is made, in particular, between mobile and trackside subsystems, with mobile subsystems including, for example, rolling stock and on-board control-command and signalling. The mobile subsystem in particular is examined in more detail below.

The modalities for approval of trackside or on-board railway products derive mainly from the Implementing Regulation (EU) 2018/545 and Directive (EU) 2016/797 on the interoperability of railway systems in the EU (TSI). The latter prescribes requirements for all parts of the railway system and the form of proof of conformity. The approval of mobile subsystems in the railway sector is largely governed by European law and is based on Directive (EU) 2016/797 on the interoperability of the railway system in the European Union. Prior to commissioning a mobile subsystem – ultimately a train – an authorization for placing into service is required. Applications for this have to be submitted to the European Railway Agency (ERA) (if the subsystem is to be used in several member states). Like other directives in the area of product safety, the Directive (EU) 2016/797 contains essential requirements that subsystems have to fulfil (Annex III to Directive (EU) 2016/797). These include, in particular, vehicle safety (Annex III No. 2.4.1 to Directive (EU) 2016/797).

The detailed procedure for verifying compliance with safety standards and applying for an authorization for placing into service is governed by the Implementing Regulation (EU) 2018/545 establishing practical arrangements for the railway vehicle authorisation and railway vehicle type authorisation process pursuant to Directive (EU) 2016/797. In order to obtain an authorization for placing into service, compliance with the essential requirements has to be demonstrated in accordance with Art. 15 of Directive (EU) 2016/797. In addition, according to Art. 13 (3) Implementing Regulation (EU) 2018/545, proof of safety and proof of implementation of a risk management process according to Regulation (EU) No. 402/2013 (for aspects not covered by TSI or national regulations) are required; it is necessary to demonstrate that the various components have been safely integrated.

The aforementioned conformity assessments are to be carried out by third parties, designated bodies or notified bodies. Among other things, these establish the proof of safety within the framework of the conformity assessment for the subsystem. The objective of this proof of safety or safety assessment is to demonstrate that a system is appropriately safe for its intended use (e.g. obstacle detection when driving on different terrain), i.e. that the subsystem does not pose a hazard. The proof of safety shows and justifies how the system fulfils the specified safety requirements. It contains proofs of quality and safety management as well as proofs of technical safety. The test is conducted in accordance with the requirements of Annex IV to Directive (EU) 2016/797, which regulates, in particular, the test steps, the content of the technical dossier attached to the test certificate

and the test certificate as such. For this purpose, the individual components have to comply with the specific product safety standards from the respective technical specifications for interoperability (TSI) and have the required functional safety.

The basis for verifying this functional safety is formed by the railway-specific standards DIN EN 50126, 50128, 50129 and the more general standards, e.g. IEC 61508, DIN EN 62061, ISO 26262 and, where applicable, DIN EN 62443, provided they are relevant to the respective subsystem. The general requirements for the proof of safety are derived from DIN EN 50129. This is part of the documentation that has to be submitted to the supervisory authority for approval. The structure of the proof of safety is defined in Sect. 7 of DIN EN 50129. The proof of safety also includes proof of proper function. For software applications, the requirements of DIN EN 50128 should also be taken into account. The "technical" proof is located in the section "Technical safety report".

In particular, the railway-specific safety standards do not provide detailed specifications for dealing with AI applications; notably, they lack references to system definitions (e.g. purpose, components, and system boundaries), conditions of use, identifications of hazards, the definition of risk acceptance criteria and even safety requirements.

Where AI is addressed, e.g. in the context of software development or data transmission (DIN EN 50128, informative Annex D.1), possible areas of application are only mentioned without elaborating on its specific characteristics. An AI-controlled system, therefore, has to be able to meet the same functional safety requirements in terms of reliability, applicability, maintenance and safety (RAMS) as any other railway system, while minimizing the effects of malfunctions in order to provide the required basic safety and reliability requirements. To this end, a risk management procedure in accordance with the Implementing Regulation (EU) No. 402/2013 has to be conducted and documented in order to demonstrate the safety of the components. In this context, the definition of tolerable hazard rates (THR) is also part of the safety requirement, which is why their compliance also has to be demonstrated in the proof of safety. The tolerable hazard rate in this regard is a measure that represents an acceptable failure rate per hour depending on the severity of the effects of a failure and corresponds to the safety integrity level (SIL). Safety integrity levels are divided into different levels from 0 to 4. For example, the THR of an application for which SIL 4 is intended lies at 10^{-9}.

At this point, a real problem arises concerning AI applications, because it needs to be described and explained that all components are able to fulfil the specifications of the system requirement and the specifications of the safety requirement as well. For instance, it has to be demonstrated that a technical requirement or system requirement can be implemented as reliably as if it would be by humans or can be as safe as an already implemented technology – e.g. obstacle detection by machine learning instead of driver/sensor technology.

Therefore, it is no longer sufficient to determine input and output, but the processes of the AI should be explained in detail ("how does the AI comes to its results–how does it determine an obstacle", "in what way are its conclusions made").

3.2 AI-Act

The draft AI-Act, which is intended to create the legal framework for the development, use and dissemination of trustworthy AI systems, also fits into the systematics of the NLF (DIN e.V., DKE, 2021).

The AI-Act assumes functions equivalent to those of a product directive within the NLF – similar to the Machinery Directive (Directive 2006/42/EC). It implements the risk-based regulatory approach known from the NLF. While AI systems with "unacceptable risk" are to be prohibited (Art 5 No. 1 AI-Act), high-risk AI systems will have to meet mandatory requirements prior to being placed on the market. This includes, in particular, the obligation to submit to a conformity assessment prior to being placed on the market or put into service (Art. 19 (1) AI-Act).

According to Art. 3 No. 1 AI-Act, the material scope of the regulation is broad. It includes all AI systems that use the technologies listed in Annex I. Basically, any software that uses technologies such as machine learning approaches, logic- and knowledge-based approaches, or statistical approaches. The personal scope is derived from Art. 2 (1) AI-Act and includes providers and users. The provider is any "natural or legal person, public authority, institution or other body that develops an AI system or has it developed with the aim of placing it on the market or putting it into operation […]" (Art. 3 No. 2 AI-Act). The user is defined as "a natural or legal person, public authority, agency or other entity that uses an AI system under its own responsibility, unless the AI system is used in the course of a personal and non-professional activity" (Art. 3 No. 4 AI-Act). To the extent that these actors and/or the AI systems they use are related to the European Market, the AI-Act is applicable (Wiebe, 2022).

As mentioned above, the AI-Act takes a risk-based approach. Therefore, the AI systems used are classified into different risk categories. The draft distinguishes between systems with unacceptable, high and low/minimal risk. Systems with unacceptable risk are prohibited from being used (Art. 5 AI-Act). For AI systems with high-risk – which constitute the core of the AI-Act – the act specifies a set of mandatory requirements before AI systems are allowed to be placed on the market (Art. 8 AI-Act). In contrast, there are hardly any restrictions for systems with only low risk.

High-risk AI systems are specified in Art. 6 AI-Act. On the one hand, these are AI systems that perform a safety function in other products (Art. 6 (1) AI-Act) and, on the other hand, AI applications that are explicitly listed in Annex III of the AI-Act (Art. 6 (2) AI-Act) as shown in Fig. 3. In the context of AI systems according to Art. 6 (1) AI-Act, the requirements specified in the regulation have to be fulfilled in a cumulative manner. Thus, the AI system first has to be used as a safety component in a product that is covered by one of the EU regulations listed in Annex II AI-Act (Art. 6 (1) (a) AI-Act). In addition, the product or the integrated AI system as a safety component is required to be subject to a conformity assessment by a third party (i.e. a notified body) in accordance with one of the harmonization regulations listed in Annex II. This requirement for an audit raises a variety of issues. The internal audits require a high level of trust in the provider of the AI system (Mökander et al., 2022), while the external audits required by the railway-specific regulations for safety components are currently not available.

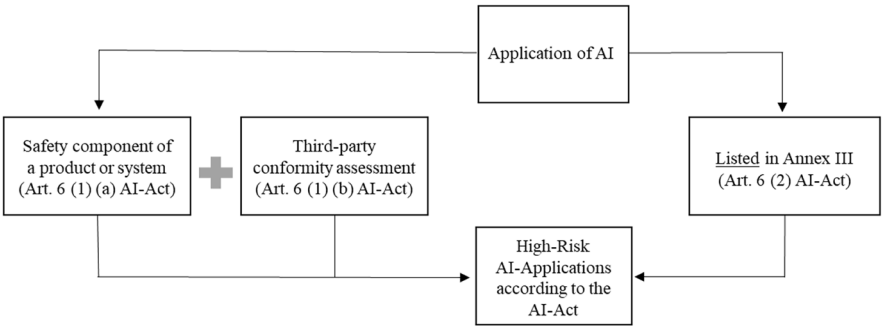

Fig. 3 Requirements high-risk AI systems

In contrast, the list in Annex III of the AI-Act – to which Art. 6 (2) AI-Act refers and which specifies AI systems that are part of the category of high-risk AI systems in any case – is kept more general. The list includes specific application scenarios, such as biometric identification of users (Annex III No. 1), management of critical infrastructure (Annex III No. 2) or law enforcement (Annex III No. 6). Under certain risk-oriented conditions, the European Commission may add further AI application scenarios in the future (Art. 7 AI-Act); therefore the list is dynamic.

Looking at the railway applications described above in the area of control and command, such as the object recognition (in Sect. 2.2.1) and FOS for position determination (in Sect. 2.2.2), it can be seen that they do not fit into the scenarios of Art. 6 (2), Annex III AI-Act, although this assessment could of course be changed if the list as such or the design of AI applications described above were changed.

Provided that the respective systems are covered by one of the harmonization directives in Annex II as a safety component or as part of a safety component and these directives provide for third-party conformity assessment, the provisions for high-risk AI systems may continue to be applied (Art. 6 (1) AI Act). In Annex II, the main directives that can be considered for rail-based systems are Directive (EU) 2016/797 on the interoperability of the rail system in the European Union and the Machinery Directive 2006/42/EC. If the AI applications are to be subsumed under these directives and a third-party conformity assessment needs to be carried out, the AI Act applies (Bomhard & Merkle, 2021). However, this results in the contradiction that Annex II of the AI-Act, divided into Parts A and B, may not be applied in its entirety. Annex II Section B corresponds to the list of harmonizing legal acts exempted from the application of the AI-Act, with the exception of Art. 84 AI-Act, according to its Art. 2 (2). (a)–(g). This exception includes Directive (EU) 2016/797, which is particularly relevant for the railway sector. The additional requirements of the AI-Act would therefore only be applicable to railway applications that (also) fall under another harmonization directive, such as the Machinery Directive.

Accordingly, the finalization of the AI-Act must be awaited in order to see whether the legislator will resolve this contradiction and possibly leave the previous railway-specific regulations as sufficient. Or, whether the harmonization provisions that have been excluded to this extent should be included in the horizontal

functioning of the AI Act, which would probably also benefit the standardization of the artificial intelligence systems used. Either way, these would have to be subjected to a conformity assessment, which could then of course be standardized. This would avoid divergence of technical requirements in terms of design and conformity assessment between the railway and other sectors.

The requirements of the AI-Act also include the obligation to conduct a conformity assessment of a high-risk AI system prior to placing it on the market or putting it into operation (Art. 19 (1) AI-Act). The design of the conformity assessment for high-risk AI systems is standardized in Art. 43 AI Act and specified in Annexes VI and VII of the AI Act. The decisive factor is whether the AI system used is a high-risk AI system according to Art. 6 (1) or Art. 6 (2) AI-Act. In order to assess the conformity of a high-risk system with an intended application from the list of Annex III to the AI-Act, it is first necessary to determine which number of Annex III is relevant. Thus, a distinction has to be made between high-risk AI systems according to Annex III No. 1 AI-Act, i.e. those serving the biometric identification of natural persons, and those according to Annex III No. 2–8 AI-Act, i.e. those that pursue another intended application from the list of Annex III.

To the extent that a high-risk AI system according to Annex III No. 1 AI-Act exists, a distinction is required again as to whether it was established in accordance with the harmonized standards according to Art. 40 AI-Act or common specifications according to Art. 41 AI-Act, or whether the harmonized standards were not applied or were applied only partially, or whether no corresponding standards exist. If the corresponding standards or specifications are fulfilled, an internal conformity assessment procedure according to Annex VI can be conducted; otherwise, a conformity assessment procedure according to Annex VII has to be carried out.

High-risk AI systems with an intended use according to Annex III No. 2–8, conversely, always have to complete a conformity assessment procedure according to Annex VI (Art. 43 II AI-Act). In the context of high-risk AI systems, as defined in Art. 6 (1) AI-Act, which are covered by Annex II Section A AI-Act, a conformity assessment has to be carried out according to the relevant provisions of the respective legal act pursuant to Art. 43 (3) AI-Act. The conformity assessment has to comply with the essential requirements of the Chapter 2 AI-Act. In addition, No. 4.3 to 4.5 and 4.5 (6) from Annex VII to the AI-Act are to be considered.

However, there is no corresponding regulation regarding Section B of Annex III AI-Act, which in turn is a consequence of the incomplete regulation of the exemption areas pursuant to Art. 2 (2) AI-Act. If the exemption from the AI-Act pursuant to Art. 2 (2) (e) AI-Act was deleted and Art 43 (3) AI-Act was to be extended to Section B of Annex III AI-Act, this would also cover the Interoperability Directive (EU) 2016/797. In that case, AI applications from the railway sector that have a safety function (including the discussed AI applications) would also be covered by the essential requirements of the AI-Act. However – it should be noted – the approach of the EU legislator is currently not predictable.

With regard to the safety requirements, the existing regulatory framework for product safety remains essentially unchanged. The AI-Act does not define these safety requirements directly, but refers to existing NLF ordinances and directives in its annexes. The AI-Act adopts the various safety requirements from the regulations

and directives by superimposing them – figuratively speaking – horizontally on the vertically arranged pillars of European product safety law. According to recital 61 of the AI-Act, the harmonized standards, which provide technical specifications for the NLF regulations and directives, are to continue to apply. If these harmonized standards are complied with, it is presumed even under the current conception of the AI-Act that the respective high-risk AI system complies not only with the sector-specific product safety law but also with the requirements of the AI-Act (Art. 40 AI-Act). The AI-Act thus incorporates the existing product-specific requirements for harmonized products and combines them with the AI-specific aspects. It clarifies that the essential requirements for high-risk AI systems continue to derive from NLF at least initially (Art. 24, 43 (3) AI-Act). In addition, on a broader level, there are the special AI-specific requirements, such as non-discrimination of training data (Art. 10 (3) AI-Act), the requirement of sufficient transparency (Art. 13 (1) AI-Act), or ethical balance – as an aspect whose guidance by norms and standards (e.g. criticality approach (Wahlster & Winterhalter, 2020)) – is currently only being explored. According to the conception of the AI-Act, it is not intended to elaborate these AI-specific requirements in the AI-Act as such; these essential requirements will also be reflected in the respective sector-specific EU-product-directives in the future.

This approach of the European legislator is appropriate since AI as a tool can ultimately only be significant in its respective product context. In this respect, AI or an AI system in its respective application context has to continue to meet the existing sector-specific functional safety requirements, which also include the railway-specific requirements.

4 Conformity Assessment

Conformity assessments according to the AI-Act and according to the railway-specific regulations ultimately lead to the application of technical standards for design and conformity assessment.

If safety-relevant railway applications contain AI systems, these will have to be subjected to conformity assessments, which in principle have to be based on the harmonized standards of the railway sector. In addition, technical standards could be applied that address the safe development of AI applications in accordance with the basic requirements of the AI-Act. However, these technical standards for AI development and, above all, for proving the functionality of AI, do currently not exist.

Another problem is that the relevant technical standards from the railway sector do not yet consider the application of AI but still require specific validation and verification of system functions for safe system development and proof of safety. Due to the extremely complex nature of AI systems, commonly referred to as “black boxes”, especially in the approaches currently being pursued based on “deep learning” through “convolutional neural networks”, the processes remain opaque (Burell, 2016). The specific processes of decision-making are currently almost impossible to understand.

This already leads to problems in advance of the actual proof of safety, namely during system development in accordance with DIN EN 50126. Here, the proof of the proper functional behaviour would already have to be provided. While the description of the

system architecture and the interface definition still appear to be without problematic, the difficulties start with the proof of the fulfilment of the system requirement specification, i.e. the functional operating requirements. This is due to the fact that, on the one hand, valid and sufficient test and test development environments have to be ensured, which becomes difficult in the case of the AI applications presented here, if – as in the case of obstacle detection – extensive training and test data sets become necessary. The situation is similar in the case of FOS, which initially requires an extensive setup.

Considering the automation efforts also in the railway sector and the implementation efforts for AI-based systems, machine learning in particular is rather a long-term process that requires extensive experimental application of this system prior to starting the process of its certification. On the other, there are significant obstacles, especially in the context of system validation. In the scope of this system validation, a proof would have to be provided that the AI system correctly interprets the requirements or learning content and that the interpretation requirements are completely and correctly fulfilled as well.

Analytical proof of the interpretation of what has been learned is probably not possible at this stage. While there are initial techniques such as LIME (Ribeiro et al., 2016) or Tree Regularization (Wu et al., 2017) that can show which part of the input affects the network's decision in one direction or another, this is not yet sufficient in order to fully validate the system.

5 Conclusion

The introduction of AI-based obstacle detection and classification, as well as AI-based determination of the train position, is problematic when meeting the requirements for safety standards. However, this is required if safety functions are to be combined with these new technical possibilities, which is precisely what is being aimed for, especially with regard to autonomous driving. In the case of the previously mentioned examples, it is currently not possible to comply with the conformity assessment requirements for proof of safety. There are currently no simple solutions to solve the problems described. Most importantly, the decision-making of the fundamental AI system cannot be understood properly from an analytical point of view. The research field of "explainable AI" does attempt to provide solutions to make the decision-making process of neural networks and other AI applications comprehensible and understandable. However, this is not an option for timely implementation of the desired AI applications; it just takes too long.

Thus, as far as that the black-box problem has not yet been solved and the high complexity of AI systems is not yet transparent, another solution has to be found for deployments in experimental environments. For example, the risk-based approach from the AI-Act could be pursued in standardization. If system development and proof of safety did not have to be carried out analytically but could instead be based on the results of extensive test series, for example, development in compliance with the standard could already be achievable. The relevant standards, in particular DIN EN 50126, 50128 and 50129, would have to be adapted accordingly.

The usual approach to certification is to achieve the required SIL. However, certification of systems incorporating AI can be problematic, especially for SIL 4, and some even recommend not providing for using AI when a level higher than SIL1 is required. An alternative approach could be to adapt the relevant standards in a way that specific validation of the system function is no longer necessary, and to replace this instead with (virtual) test series – for example, in the context of a digital twin. However, whether such an approach should actually be pursued is worth discussing in terms of the extensive use and high risk that safety functions encounter in the railway sector.

Instead, a reasonable approach to solve this problem could be to deploy AI-based systems with responsibility of the driver (in terms of obstacle and object detection) or the control centre (in terms of FOS) in ATO Level 3 (attended train operation). Using this system on a large-scale could allow adequate training of the AI system for obstacle classification (obstacle) and validation of speed as well as the accuracy of transmission rates (FOS), raising the safety level to that required to achieve SIL 4.

In this sense, both the legal regulations and the corresponding technical standards are likely to be consistent in their principles with the latest developments in the software field.

Author Contribution On behalf of all authors, the corresponding author states that both authors contributed equally to the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Competing Interests The authors declare no competing interests.

References

Bittner, J., Debowski, N., Lorenz, M., Raber, H., Stege, H., & Teille, K. (2021). Recht und Ethik bei der Entwicklung von Künstlicher Intelligenz für die Mobilität. *Neue Zeitschrift Für Verkehrsrecht, 10*, 505–524.

Bomhard, D., & Merkle, M. (2021). Europäische KI-Verordnung – Der aktuelle Kommissionsentwurf und praktische Auswirkungen. *Recht Digital, 6*, 276–283.

Braband, J. (2021). Künstliche Intelligenz – mit Sicherheit? *Deine Bahn, 4*, 30–36.

Braband, J., & Schäbe, H. (2020). On safety assesment of artificial intelligence. *Dependability*. https://doi.org/10.21683/1729-2646-2020-20-4-25-34

Burell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*. https://doi.org/10.1177/2053951715622512

Dagvasumberel, A., Myagmardulam, B., Myagmar, B., & Nakayama, T. (2021). *Railway near-miss occurrence detection and risk estimation system with data from camera using deep learning*. 5th International Conference on Imaging, Signal Processing and Communications (ICISPC). https://doi.org/10.1109/ICISPC53419.2021.00023

DIN e.V. DKE. (2021). *Position paper on the EU "Artificial Intelligence Act"*. Retrieved February 28, 2022, from https://www.din.de/resource/blob/800324/c50ed443e81c47f8860b3f5c2b3b0742/21-06-din-dke-position-paper-artificial-intelligence-act-data.pdf

Ensthaler, J., & Gesmann-Nuissl, D. (2006). *Gestaltungsspielräume für staatliche Aufsichtssysteme angesichts des Vorsorgeprinzips im deutschen Produktsicherheitsrecht und den Entwicklungen im europäischen Raum*. transfer-Verlag.

Ensthaler, J., Gesmann-Nuissl, D., & Müller, S. (2012). *Technikrecht*. Springer Nature.

He, D., Zou, Z., Chen, Y., Liu, B., Yao, X., & Shan, S. (2021). Obstacle detection of rail transit based on deep learning. *Measurement*. https://doi.org/10.1016/j.measurement.2021.109241

Kapoor, A., & Klindt, T. (2008). New legislative framework" im EU-Produktsicherheitsrecht – Neue Marktüberwachung in Europa? *Europäische Zeitschrift Für Wirtschaftsrecht, 19*, 649–655.

Kaulartz, M., & Braegelmann, T. (2020). *Rechtshandbuch. Artificial Intelligence und Machine Learning*. Vahlen.

Kowarik, S., Hussels, M.-T., Chruscicki, S., Münzenberger, S., Lämmerhirt, A., Pohl, P., & Schubert, M. (2020). Fiber optic train monitoring with distributed acoustic sensing: Conventional and neural network data analysis. *Sensors, 20*(2), 450. https://doi.org/10.3390/s20020450

Mockel, S., & Scherer, F. (2003). Multi-sensor obstacle detection on railway tracks. *Intelligent Vehicles Symposium* (pp. 42–46). IEEE. https://doi.org/10.1109/IVS.2003.1212880

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*. https://doi.org/10.1007/s11023-021-09577-4

Osborne, C. (2021). *The European Commission's Artificial Intelligence Act highlights the need for an effective AI assurance ecosystem*. Centre for Data Ethics and Innovation Blog. Retrieved June 9, 2022, from https://cdei.blog.gov.uk/2021/05/11/the-european-commissions-artificial-intelligence-act-highlights-the-need-for-an-effective-ai-assurance-ecosystem/

Reinhold, T., & Kasperkovitz, G. (2013). Eisenbahn in Deutschland 2025 – Zukunftsperspektiven für Mobilität und Logistik. *Zukunftsforschung im Praxistest* (pp. 299–319). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-531-19837-8_13

Ribeiro, M., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Retrieved June 9, 2022, from https://arxiv.org/abs/1602.04938

Ristić-Durrant, D., Abdul Haseeb, M., Banić, M., Stamenković, D., Simonović, M., & Nikolić, D. (2021). SMART on-board multi-sensor obstacle detection system for improvement of rail transport safety. *Journal of Rail and Rapid Transit*. https://doi.org/10.1177/09544097211032738

Russell, S., & Norvig, P. (2012). *Künstliche Intelligenz*. Pearson.

Vidovic, I., & Landgraf, M. (2018). *Fibre optic sensing as innovative tool for evaluating railway track condition?* International Conference on Smart Infrastructure and Construction 2018 (ICSIC). https://doi.org/10.1680/icsic.64669.107

Vidovic, I., & Maschnig, S. (2020). Optical fibres for condition monitoring of railway infrastructure–encouraging data source or errant effort? *Applied Sciences*. https://doi.org/10.3390/app10176016

Wahlster, W., & Winterhalter, C. (2020). *Deutsche Normungsroadmap Künstliche Intelligenz*. Retrieved February 28, 2022, from https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Weichselbaum, J., Zinner, C., Gebauer, O., & Pree, W. (2013). Accurate 3D-vision-based obstacle detection for an autonomous train. *Computers in Industry*. https://doi.org/10.1016/j.compind.2013.03.015

Wiebe, A. (2022). *Produktsicherheitsrechtliche Betrachtung des Vorschlags für eine KI-Verordnung* (pp. 899–906). Betriebs-Berater.

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., Doshi-Velez, F. (2017). *Beyond sparsity: Tree regularization of deep models for interpretability*. Retrieved June 9, 2022, from https://arxiv.org/abs/1711.06178

Yu, M., Yang, P., & Wei, S. (2018). Railway obstacle detection algorithm using neural network. *AIP Conference Proceedings*. https://doi.org/10.1063/1.5039091