Entwicklung vertrauenswürdiger Algorithmen auf Basis Künstlicher Intelligenz

The development of trustworthy algorithms based on artificial intelligence

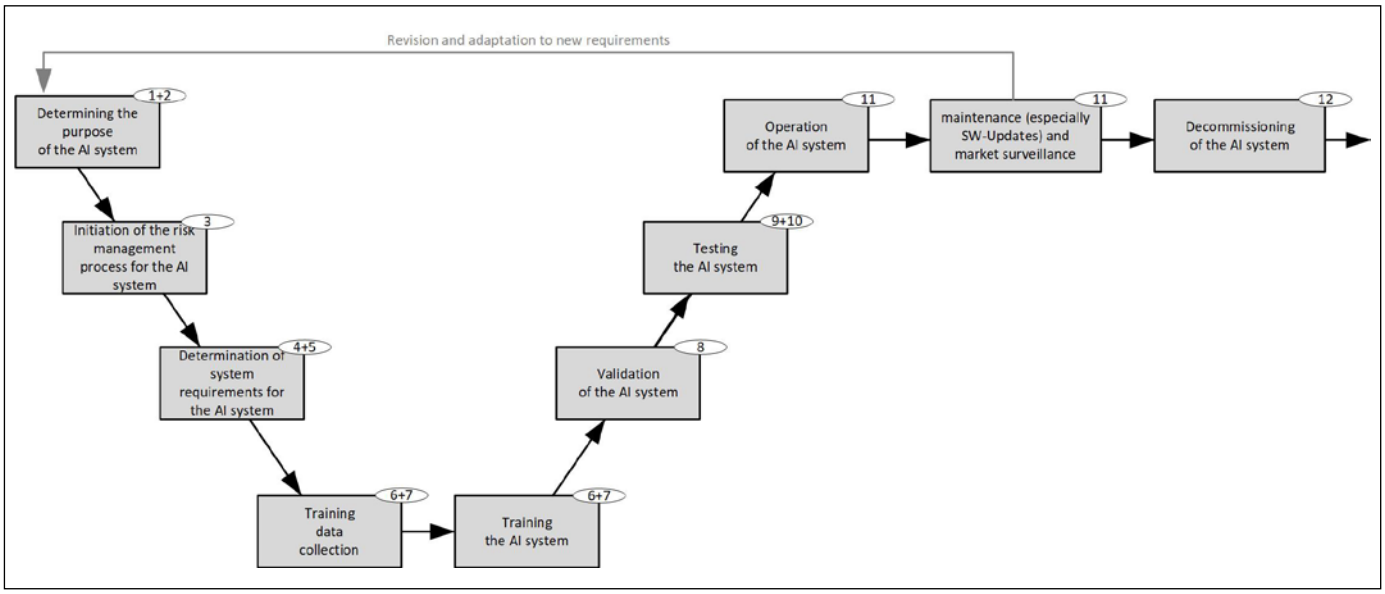

Bild 1: Lebenszyklus eines KI-Systems nach [6] und seine Abbildung auf den Lebenszyklus für Bahnanwendungen nach [4]
Fig. 1: The lifecycle of an AI system according to [6] and its depiction on the lifecycle for railway applications according to [4]

Lars Schnieder

Vertrauenswürdige Systeme der Künstlichen Intelligenz (KI-Systeme) werden zukünftig die Grundlage für einen wettbewerbsfähigen Schienenverkehr sein. Ein KI-System ist eine Software, die mit einer oder mehreren Techniken und Konzepten entwickelt worden ist und im Hinblick auf eine Reihe von Zielen, die vom Menschen festgelegt werden, Ergebnisse wie Inhalte, Vorhersagen, Empfehlungen oder Entscheidungen hervorbringen kann, die das Umfeld beeinflussen, mit dem sie interagieren. Dieser Beitrag skizziert einen Rahmen für den Entwurf, die Implementierung und die Zulassung vertrauenswürdiger KI [1].

1 Methodische Grundlagen eines KI-spezifischen Risikomanagements

Im Risikomanagement wird der Lebenszyklus des KI-Systems systematisch hinsichtlich relevanter Risiken analysiert. Das Vorgehen basiert auf zwei aufeinander aufbauenden Schritten:

1. **Modellierung des Lebenszyklus des KI-Systems:** Der Entwicklungslebenszyklus von KI-Systemen ist komplex und vollzieht sich in mehreren Schritten. Als Grundlage einer methodischen Betrachtung ist es zielführend, diesen Ablauf mit Beschreibungsmitteln zu analysieren. Durch die Semantik des Beschreibungsmittels werden bestimmte Aspekte (bspw. kausale Abhängigkeiten ein-

Trusted artificial intelligence (AI) systems will constitute the basis of competitive rail transport in the future. An AI system involves software that has been developed using one or more techniques and concepts and is capable of producing results such as content, predictions, recommendations or decisions that influence the environment with which it interacts in relation to a set of goals specified by humans. This article outlines a framework for the design, implementation and approval of trustworthy AI [1].

1 The methodological foundations of AI-specific risk management

Risk management systematically analyses an AI system's lifecycle with regard to any relevant risks. The procedure is based on two steps that build on one another:

1. **Modelling the AI system's lifecycle:** The development of an AI system's lifecycle is complex and occurs in several steps. It is useful to analyse this process using means of description as the basis for a methodical approach. The semantics of the means of description emphasise certain aspects (e.g. the causal dependencies of the individual processing steps) [2]. The representation of the lifecycle is based on the vaguely de-

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

zelner Bearbeitungsschritte) hervorgehoben [2]. Die Darstellung des Lebenszyklus basiert auf dem vage beschriebenen Vorgehen nach [3], welches auf Basis einer Literaturrecherche erweitert wurde und den in der Entwicklung von Bahnanwendungen üblichen Lebenszyklusphasen [4] gegenübergestellt wurde. Der Lebenszyklus stellt keinen linearen Prozess dar. Es kann vielmehr notwendig sein, mehrmals zu einer früheren Phase zurückzukehren. Ferner sind begleitende qualitätssichernde Maßnahmen erforderlich. Bevor mit der nächsten Phase begonnen wird, ist zu prüfen, ob alle Anforderungen der jeweils aktuellen Phase erfüllt wurden (Bild 1).

2. **Analyse des Entwurfsprozesses mittels FMEA:** Die „Failure Modes and Effects Analysis" (FMEA) ist in der Entwicklung zuverlässiger technischer Systeme bewährt [5]. Die FMEA zielt darauf ab, alle potenziellen Ausfallarten und deren Ursachen und Auswirkungen zu identifizieren, um damit unerwünschte Auswirkungen zu reduzieren. Die FMEA dokumentiert die systematische und nachvollziehbare Identifikation von Risiken. Sie begründet erforderliche Maßnahmen zur Risikobeherrschung.

2 Risikomanagement in den Phasen des KI-Lebenszyklus

Nachfolgend werden die einzelnen Lebenszyklusphasen von KI-Systemen analysiert, Risiken identifiziert und Gegenmaßnahmen motiviert.

2.1 Festlegung des Zwecks des KI-Systems

Die Verwendung, für die ein KI-System laut Anbieter bestimmt ist, ist festzulegen. Dies umfasst die besonderen Nutzungsumstände und -bedingungen. Diese Randbedingungen sind dem Nutzer in der Anwendungsdokumentation mitzuteilen [1]. Dies entspricht in der Entwicklung von Bahnanwendungen den Phasen 1 (Konzept) und 2 (Systemdefinition und betrieblicher Kontext):

- **Potenzielle Fehler und Fehlerauswirkung:** In dieser Lebenszyklusphase entstehen Fehler in den Anforderungen. Hierbei handelt es sich um mehrdeutige, missverständliche, fehlende, falsche oder inkonsistente Anforderungen. Jeder dieser Fehler äußert sich in einem Fehlverhalten des KI-Systems.
- **Fehlerursachen:** Die Ursachen fehlerhafter Anforderungen sind unterschiedlicher Natur. Ein möglicher Grund ist die Anforderungsdokumentation in natürlicher Sprache (missverständliche Anforderung). Gegebenenfalls fehlt den Entwicklern das ausreichende Wissen über das Anwendungsgebiet (fehlende Anforderung), oder sie vertrauen zu sehr ausschließlich auf ihr Fachwissen (falsche Anforderung). Möglicherweise wird der definierte Entwicklungsprozess nicht eingehalten (unvollständige Anforderung).
- **Risikobehandlung:** Diesen Fehlern kann durch die Verwendung semi-formaler Beschreibungsmittel in der Dokumentation von Anforderungen, der Durchführung von Anforderungsreviews bzw. der gezielten Einbindung von Kunden und potenziellen Nutzern in die Anforderungsanalyse begegnet werden.

2.2 Initiierung des Risikomanagementverfahrens für das KI-System

In dieser Phase erfolgt eine Ermittlung und Analyse bekannter und vorhersehbarer Risiken, die von jedem KI-System ausgehen. Risiken, die entstehen können, wenn das KI-System entsprechend seiner Zweckbestimmung oder im Rahmen einer vernünftigerweise vorhersehbaren Fehlanwendung verwendet wird, sind abzuschätzen und zu bewerten. Außerdem erfolgt eine Bewertung anderer möglicherweise auftretender Risiken auf Basis von Felderfahrungen. Dies

scribed procedure according to [3], which has been expanded on the basis of literature research and compared to the lifecycle phases commonly used in the development of railway applications [4]. The lifecycle does not represent a linear process. Rather, it may be necessary to repeatedly return to an earlier phase. Furthermore, accompanying quality assurance measures are also necessary. It is necessary to check whether all the requirements in the given phase have been met before proceeding to the next phase (fig. 1).

2. **An analysis of the design process using FMEA:** Failure Modes and Effects Analysis (FMEA) is well established in the development of reliable technical systems [5]. An FMEA aims to identify all the potential failure modes and their causes and effects in order to reduce any undesired effects. The FMEA documents the systematic and traceable risk identification. It initiates the necessary measures for risk alleviation.

2 Risk management in the phases of the AI lifecycle

The following sections analyse the individual lifecycle phases of AI systems and identifys and initiate countermeasures.

2.1 Determining the AI system's purpose

The use, for which an AI system is suited according to the manufacturer, has to ascertained. This includes the particular circumstances and conditions of its use. These boundary conditions have to be communicated to the user in the instructions for use [1]. This corresponds to phases 1 (concept) and 2 (system definition and operating context) in the development of railway applications:

- **Potential failures and their impacts:** Any errors in the requirements occur in this lifecycle phase. This involves ambiguous, misleading, missing, incorrect or inconsistent requirements. Each of these failures manifests itself in a misbehaviour in the AI system.
- **The causes of these failures:** The causes of incorrect requirements differ. One possible reason is the requirement documentation in a natural language (a misleading requirement). Possibly the developers lack knowledge about the application area (a missing requirement) or overly trust their expertise (an incorrect requirement). Possibly the defined development process has not been followed (an incomplete requirement).
- **Risk treatment:** These failures can be avoided by using semi-formal means of description in the requirement documentation, by conducting requirement reviews or by means of the targeted involvement of customers and potential users in the requirement analysis.

2.2 Initiating the risk management process in the AI system

This phase involves the identification and analysis of any known and foreseeable risks posed by AI systems. Any risks that could arise when the AI system is used for its intended use or within the context of a reasonably foreseeable misuse will be assessed and evaluated. In addition, an assessment of any other risks that may arise will also be made based on field experience. This corresponds to phase 3 (risk analysis and assessment) in the development of railway applications:

- **Potential failures and their impacts:** The European Union Commission has proposed a risk-based approach to AI development [6]. Therefore, a missing or incorrect risk assessment of an AI system used for its intended purpose consti-

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

entspricht in der Entwicklung von Bahnanwendungen der Phase 3 (Risikoanalyse und -beurteilung), siehe hierzu auch die Gegenüberstellung in Bild 1:

- **Potenzielle Fehler und Fehlerauswirkung:** Die Kommission der Europäischen Union hat einen risikoorientierten Ansatz für die Entwicklung von KI vorgeschlagen [6]. Daher ist eine fehlende bzw. fehlerhafte Risikoeinstufung des KI-Systems, welches gemäß seinem Einsatzzweck verwendet wird, zu Beginn des Lebenszyklus eine wesentliche Fehlerquelle. Über den bestimmungsgemäßen Einsatz des KI-Systems hinaus sind auch Risiken aus der vorhersehbaren Fehlanwendung zu betrachten, die ggf. im Rahmen der Risikoanalyse unterschätzt wurden. Die Folge aus der fehlerhaften Durchführung des Risikomanagementverfahrens ist, dass das KI-System nicht den zwingend vorgeschriebenen Anforderungen genügt. Daraus resultieren zu erwartende Fehlleistungen der KI.
- **Fehlerursachen:** Ursachen für die zuvor genannten Fehler liegen zum einen in mangelnder Erfahrung in der Anwendung des Risikomanagementverfahrens bzw. seiner unvollständigen Anwendung. Zum anderen liegt eine Ursache in der unvollständigen Identifikation möglicher Ursachen vorhersehbarer Fehlanwendungen.
- **Risikobehandlung:** Um die genannten Fehler zu vermeiden, wird zukünftig die Anwendung eines Risikomanagementverfahrens verbindlich vorgegeben. Darüber hinaus ist ein Plan zur Marktbeobachtung des KI-Systems eine zwingende Voraussetzung zum Inverkehrbringen des KI-Systems. Dieser Plan sowie die stringente Umsetzung des Risikomanagementverfahrens unterliegen einer unabhängigen Konformitätsbewertung. In Bezug auf eine möglichst vollständige Identifikation von Fehlgebrauchsszenarien sei auf systematische Ansätze der Automobilindustrie verwiesen (Anhang E in [7]).

2.3 Festlegung von Systemanforderungen für das KI-System

In dieser Phase erfolgt eine ausführliche Beschreibung der konkreten Umsetzung der aus der Risikobewertung abgeleiteten Anforderungen an das KI-System. Ziel ist eine weitest mögliche Beseitigung oder Verringerung der Risiken durch eine geeignete Konzeption und Entwicklung des KI-Systems. Gegebenenfalls sind angemessene Minderungs- und Kontrollmaßnahmen für nicht auszuschließende Risiken umzusetzen. Nicht beherrschbare, aber generell akzeptierbare Restrisiken erfordern eine angemessene Information und Schulung der Nutzer. Dies entspricht in der Entwicklung von Bahnanwendungen den Phasen 4 (Festlegung von Systemanforderungen) und 5 (Architektur und Aufteilung von Systemanforderungen):

- **Potenzielle Fehler und Fehlerauswirkung:** In dieser Phase liegen zwei grundsätzliche Fehlerkategorien vor. Erstens liegen potenzielle Fehler darin begründet, dass die aus der Entwicklung resultierenden Restrisiken im Betrieb das Maß eines akzeptierten Risikos übersteigen. Zweitens können die in der Entwicklung für den Nutzer als beherrschbar angenommenen Restrisiken sich in der Praxis als nicht beherrschbar herausstellen. Beide Fälle führen zu im Betrieb nicht akzeptablen Risiken.
- **Fehlerursachen:** Die Ursachen für den ersten potenziellen Fehler liegen möglicherweise in der bewussten Umgehung der Anforderungen des Richtlinienentwurfs der Europäischen Kommission [6] durch die Hersteller oder mangelnde Expertise der Hersteller im Umgang mit neuartigen Ansätzen in KI-Systemen. Gegebenenfalls wird der Nutzer auch nicht vollständig und korrekt über bestehende Restrisiken aufgeklärt, da diese Anforderungen unzureichend an die technische Dokumentation weitergeleitet wurden oder selbst unvollständig sind (siehe oben). Die Ursachen für den

tutes a major source of failure at the beginning of the lifecycle. In addition to the intended use of the AI system, the risks from any foreseeable misuse must also be considered and may be underestimated in the risk analysis. The incorrect implementation of the risk management procedure results in the AI system not meeting the mandatory requirements. This then leads to failures that have to expected in the AI during operations.

- **The causes of these failures:** The causes of the previously mentioned failures lie firstly in a lack of experience in the application of the risk management procedure or the failure to apply it completely. Secondly, one cause involves the incomplete identification of possible causes of foreseeable misuse.
- **Risk treatment:** The application of a risk management procedure will be mandatory in the future in order to avoid the aforementioned failures. In addition, a market surveillance plan for the AI system is also a mandatory requirement for placing an AI system on the market. This plan and the stringent implementation of the risk management procedure are subject to an independent conformity assessment. Reference is made to the automotive industry's systematic approaches (Annex E in [7]) with regard to the most complete possible identification of misuse scenarios.

2.3 Determining the system requirements for the AI system

This phase involves a detailed description of the concrete implementation of the requirements for the AI system derived from the risk assessment. The aim is to eliminate or reduce the risks as far as possible through the appropriate design and development of the AI system. Where appropriate, adequate mitigation and control measures will be implemented for any risks that cannot be eliminated. Residual risks that cannot be controlled, but are generally acceptable, require appropriate information and user training. This corresponds to phases 4 (definition of the system requirements) and 5 (architecture and partitioning of system requirements) in the development of railway applications:

- **Potential failures and their impacts:** Two basic categories of failures are present in this phase. First, potential failures are due to the fact that the residual risks resulting from development exceed the accepted risk level during operations. Secondly, the residual risks that are assumed to be manageable for the user during development may turn out to be unmanageable in practice. Both cases can lead to risks that are unacceptable in operations.
- **The causes of these failures:** The causes of the first potential failure may lie in the deliberate circumvention of the requirements of the European Commission's draft directive [6] by manufacturers or result from a lack of expertise among manufacturers in dealing with novel approaches to AI systems. Otherwise, the user may also not have been fully and correctly informed of any existing residual risks, because these requirements have been insufficiently incorporated into the technical documentation (or are incomplete, see above). The causes for the second potential failure are rooted in a lack of field experience in dealing with AI systems.
- **Risk treatment:** An effective measure involves the implementation of an independent conformity assessment accompanying the development. This ensures that appropriate risk management is maintained at all times during the development process as a continuous iterative process with regular systematic updates. This ensures the hierarchy of the safety measure

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

zweiten potenziellen Fehler liegen in der mangelnden Felderfahrung im Umgang mit KI-Systemen begründet.

- **Risikobehandlung:** Eine wirksame Maßnahme ist die Umsetzung einer entwicklungsbegleitenden unabhängigen Konformitätsbewertung. Dies stellt sicher, dass zu jedem Zeitpunkt während des Entwicklungsprozesses ein angemessenes Risikomanagement als kontinuierlicher iterativer Prozess mit regelmäßiger systematischer Aktualisierung gepflegt wird. Dies gewährleistet, dass die Rangordnung der Umsetzung von Sicherheitsmaßnahmen – an deren Ende die Instruktion des Nutzers steht – korrekt von den Herstellern der KI-Systeme umgesetzt wurde. Bereits in dieser Lebenszyklusphase wird durch eine durchgängige Rückverfolgbarkeit von Anforderungen (Traceability) von der Risikoanalyse zur Übernahme sicherheitsbezogener Anwendungsregeln in die Anwenderdokumentation die Grundlage für die Durchgängigkeit des Risikomanagementverfahrens über den gesamten Lebenszyklus geschaffen. Darüber hinaus wird auch die systematische Erfassung von Felderfahrungen und ihr Rückfluss in die Entwicklung eine zunehmend größere Rolle spielen.

2.4 Erfassung von Trainingsdaten

Daten sind erforderlich, um das KI-System zu trainieren. Diese müssen ggf. unter Berücksichtigung datenschutzrechtlicher Belange erfasst werden. Mittels der Trainingsdaten werden beispielsweise lernbare Parameter und die Gewichte neuronaler Netze angepasst, sofern solche Verwendung finden. Dies entspricht in der Entwicklung von Bahnanwendungen den Phasen 6 (Entwurf und Implementierung) und 7 (Herstellung):

- **Potenzielle Fehler und Fehlerauswirkung:** In dieser Phase werden drei Fehlerursachen unterschieden. Erstens wirkt sich eine unzureichende Qualität der Trainingsdaten negativ auf die (Sicherheits-)Leistungsfähigkeit des KI-Systems aus. Datenqualität umfasst die Dimensionen der Relevanz, der Repräsentativität bzw. Aktualität, der Fehlerfreiheit sowie der Vollständigkeit (insbesondere hinsichtlich ihrer Tauglichkeit für den Test von Systemgrenzbedingungen). Sind die Daten unvollständig, so werden z. B. nicht alle für die Vermeidung gefährlicher Situationen relevanten Szenarien abgedeckt. Zweitens wirkt sich eine unzureichende Quantität der Trainingsdaten auf die erreichbare Güte der KI-Algorithmen aus. In der Statistik werden oft Parameter der Grundgesamtheit aufgrund einer Stichprobe geschätzt. Diese Schätzung wird mit zunehmender Größe der Stichprobe genauer. Drittens kann es zu einer Modifikation von Trainingsdaten (Datenvergiftung, Data Poisoning Attack) kommen. Dieser Angriff manipuliert den Trainingsdatensatz, um das Vorhersageverhalten eines trainierten KI-Modells zu beeinflussen. Alle genannten Fehler führen dazu, dass das KI-Modell eine nicht korrekte Vorhersage treffen wird oder dies lediglich mit einer geringen Wahrscheinlichkeit tun wird.
- **Fehlerursachen:** Ursachen für die zuvor genannten Fehler liegen zum einen in einer unzureichenden Spezifikation der Datenerhebung, einer mangelhaften Qualitätssicherung der Datenerfassung, zum anderen in einer inkonsequenten Bewirtschaftung der Daten über den Lebenszyklus des KI-Systems begründet.
- **Risikobehandlung:** Die Absicherung von Qualität und Quantität der Trainingsdaten erfordert ein professionelles Datenmanagement. Dieses beginnt mit einer genauen Spezifizierung der Anforderungen an die Daten. Auch sind die erfassten Datenbestände einem stringenten Konfigurationsmanagement zu unterwerfen. Hierbei werden Daten mit Zeitstempel und Versionsnummer versehen, Datenänderungen protokolliert, Quellen eindeutig zugeordnet und besondere Vorsichtsmaßnahmen bei Verwendung von Daten externer Quellen umgesetzt. Um die Datenqualität

implementation culminating in the instruction of the user – this has been correctly implemented by the manufacturers of the AI systems. The basis for the consistency of the risk management process throughout the entire lifecycle is established in this lifecycle phase by means of the continuous traceability of requirements from the risk analysis to the adoption of the safety-related application rules in the user documentation. In addition, the systematic recording of field experience and its feedback into development will also play an increasingly important role.

2.4 Training data collection

Data is required to train the AI system. Depending on the application, the data may need to be collected while taking data privacy law into consideration. This involves adjusting its learnable parameters and the weights of the neural network, if one is used. This corresponds to phases 6 (design and implementation) and 7 (production) in the development of railway applications:

- **Potential failures and their impacts:** Three causes of these failures are distinguished in this phase. First, insufficient training data quality has a negative impact on the (safety) performance of an AI system. Data quality includes the dimensions of relevance, representativeness or timeliness, freedom from errors and completeness (especially with regard to its suitability for testing system boundary conditions). If the data is incomplete, not all the scenarios relevant for the avoidance of dangerous situations will have been covered. Secondly, an insufficient amount of training data affects the achievable quality of the AI algorithms. In statistics, population parameters are often estimated based on a sample. This estimate becomes more accurate as the sample size increases. Thirdly, there can also be a modification to the training data (a data poisoning attack). This attack manipulates the training data set to influence the predictive behaviour of the trained AI model. All of the above failures will result in the AI model making an incorrect prediction or doing so with a low probability.
- **The causes of these failures:** The causes for the previously mentioned failures are partly due to the insufficient specification of the data collection, insufficient quality assurance during the data collection or inconsistent management of the data over the lifecycle of the AI system.
- **Risk treatment:** Assuring the quality and quantity of the training data requires professional data management. This begins with the precise specification of the data requirements. The collected data sets must also be subjected to stringent configuration management. This involves providing data with a time stamp and version number, logging any data changes, clearly assigning sources and implementing special precautions when using data from external sources. A regular data review for timeliness is required to maintain the data quality throughout the lifecycle. This may also include the implementation of independent and objective third-party verification of the data sets used in the development (conformity assessment). The data must also be protected against any unauthorised access by third parties [8].

2.5 Training the AI system

The learnable parameters and the weights of the AI system are adjusted in this phase. This corresponds to phases 6 (design and implementation) and 7 (production) in the development

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

über den gesamten Lebenszyklus aufrecht zu erhalten, ist eine regelmäßige Überprüfung der Daten auf Aktualität und Vollständigkeit erforderlich. Dies kann auch die Umsetzung einer unabhängigen und objektiven Überprüfung der in der Entwicklung verwendeten Datensätze durch Dritte umfassen (Konformitätsbewertung). Auch sind die Daten gegen unberechtigten Zugriff Dritter zu schützen [8].

2.5 Trainieren des KI-Systems

In dieser Phase erfolgt eine Anpassung der lernbaren Parameter und der Gewichte des KI-Systems. Dies entspricht in der Entwicklung von Bahnanwendungen den Phasen 6 (Entwurf und Implementierung) und 7 (Herstellung). Aufgrund der Bedeutung der Erfassung von Trainingsdaten und des Trainings des KI-Systems werden diese Phasen hier gesondert beschrieben:

- **Potenzielle Fehler und Fehlerauswirkung:** Ein Fehler ist die Überanpassung des KI-Systems an den Trainingsdatensatz (Overfitting). In diesem Fall lernt das KI-Modell die Details und das Rauschen in den Trainingsdaten in einem Maße, dass sich dies negativ auf die Leistung des Modells bei neuen Daten auswirkt. Das Rauschen oder die zufälligen Schwankungen in den Trainingsdaten werden als Konzepte gelernt. Die eingelernten Konzepte sind nicht auf neue Daten anwendbar, und die Verallgemeinerungsfähigkeit des Modells ist eingeschränkt. Dies hat massive Auswirkungen auf die Fähigkeit des KI-Systems, korrekte Vorhersagen zu treffen.
- **Fehlerursachen:** Es lassen sich zwei grundlegende Fehlerursachen differenzieren. Erstens ist dies eine geringe Anzahl von Beobachtungen in der Trainingsmenge im Vergleich zu den Einflussvariablen sowie eine Verzerrung (Bias) bei der Auswahl der Stichproben aus der Grundgesamtheit. Zweitens werden die Modelle zu stark trainiert.
- **Risikobehandlung:** Die erste Fehlerursache liegt in einer unzureichenden Datenqualität und -quantität begründet. Die korrespondierenden Gegenmaßnahmen wurden bereits im Zusammenhang mit den Trainingsdaten im vorherigen Abschnitt erörtert. Das übertriebene Training eines KI-Modells wird durch eine gezielte Validierung offenbart, bei der sich die fehlende Verallgemeinerungsfähigkeit des Algorithmus auf unerwartete Eingabedaten zeigt.

2.6 Validierung des KI-Systems

In dieser Phase erfolgt eine Bewertung des trainierten KI-Systems, die Abstimmung seiner nicht lernbaren Parameter sowie seines Lernprozesses. Die Durchführung der Validierung vermeidet eine Überanpassung. Der Validierungsdatensatz kann hierbei ein separater Datensatz oder Teil des Trainingsdatensatzes mit fester oder variabler Aufteilung sein. Dies entspricht in der Entwicklung von Bahnanwendungen der Phase 8 (Integration):

- **Potenzielle Fehler und Fehlerauswirkung, Fehlerursachen, Risikobehandlung:** Da in dieser Phase auf eine Teilmenge der erfassten Trainingsdaten zurückgegriffen wird, gelten die Ausführungen des Abschnitts 2.4 vollinhaltlich analog.

2.7 Test des KI-Systems

In dieser Phase erfolgt eine unabhängige Bewertung des trainierten und validierten KI-Systems, um die erwartete Leistung des KI-Systems vor Inbetriebnahme zu bestätigen. Das Testen erfolgt anhand vorab festgelegter Parameter und probabilistischer Schwellenwerte, die für die Zweckbestimmung des KI-Systems geeignet sind. Dies entspricht in der Entwicklung von Bahnanwendungen den Phasen 9 (Systemvalidierung) und 10 (Systemabnahme):

of railway applications. These phases are described separately here due to the importance of training data acquisition and the training of the AI system:

- **Potential failures and their impacts:** One failure involves the overfitting of the AI system to the training data set. In this case, the AI model learns the details and noise in the training data to such an extent that it negatively affects the model's performance with new data. The noise or random fluctuations in the training data are learned as concepts. The learnt concepts are not applicable to any new data and the generalisability of the model is limited. This has a massive impact on the ability of the AI system to make correct predictions.
- **The causes of these failures:** Two basic causes of these failures can be differentiated. The first involves a small number of observations in the training set compared to the influencing variables and the bias in the selection of the sample from the population. In the second cause, the models are overtrained.
- **Risk treatment:** The first cause of failure is due to insufficient data quality and quantity. The corresponding countermeasures have already been discussed within the context of the training data in the previous section. The overtraining of an AI model is revealed by a targeted validation, which reveals the algorithm's lack of generalisability to an unexpected input data.

2.6 Validating the AI system

The trained AI system, the tuning of its non-learning parameters and its learning process are all evaluated in this phase. The performance of validation avoids overfitting. The validation data set can be a separate data set or part of the training data set with fixed or variable partitioning. This corresponds to phase 8 (integration) in the development of railway applications:

- **Potential failures and their impacts, causes and risk treatment:** The explanations of section 2.4 apply analogously due to the fact that a subset of the collected training data is used in this phase.

2.7 Testing the AI system

An independent assessment of the trained and validated AI system is performed in this phase to confirm the expected performance of the AI system before it is put into operation. Testing is performed against pre-determined parameters and probabilistic thresholds appropriate for the intended purpose of the AI system. This corresponds to phases 9 (system validation) and 10 (system acceptance) in railway application development:

- **Potential failures and their impacts:** Since the performance of independent tests is also data based, the remarks in section 4.4 apply analogously. However, it is necessary to additionally point out that there are principle-related difficulties or limitations in demonstrably verifying the AI system when using AI systems. This leads to only limited trust in the specified properties of the AI system (acceptance).
- **The causes of these failures:** The difficulties in verifying the desired properties of AI systems have two causes. First, AI systems lack transparency and explicability because the models used in the AI system are large and complex. Also, the design and structure of AI models are principally unclear ("black box"). Secondly, the conformity assessment body may not have the necessary expertise and resources (scientific staff) for its tasks.

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

- **Potenzielle Fehler und Fehlerauswirkung:** Da auch die Durchführung unabhängiger Tests auf Daten beruht, gelten die Ausführungen des Abschnittes 2.4 analog. Allerdings wird zusätzlich darauf hingewiesen, dass es beim Einsatz von KI-Systemen prinzipbedingte Schwierigkeiten bzw. Begrenzungen darin gibt, das KI-System nachweislich zu verifizieren. Dies führt zu einem nur eingeschränkten Vertrauen in die spezifizierten Eigenschaften des KI-Systems (Akzeptanz).
- **Fehlerursachen:** Die Schwierigkeit der nachweislichen Verifikation der gewünschten Eigenschaften von KI-Systemen hat zwei Ursachen. Erstens mangelt es KI-Systemen an Transparenz und Erklärbarkeit, da die im KI-System verwendeten Modelle groß und komplex sind. Auch sind Aufbau und Struktur von KI-Modellen prinzipbedingt unklar („Black Box"). Zweitens ist es möglich, dass die Konformitätsbewertungsstelle nicht über die erforderliche Expertise und Ressourcen (wissenschaftliches Personal) für diese Aufgaben verfügt.
- **Risikobehandlung:** Zur Vermeidung der Fehlerursachen sind verschiedene Ansätze denkbar. Mit dem Ziel einer Verbesserung von Transparenz und Erklärbarkeit von KI-Systemen können im Entwurf und in der Entwicklung einfache und transparente Modelle mit reduzierter Parameteranzahl verwendet werden. Mit dem Ziel der Absicherung der Qualitätssicherungskette wird diese einer unabhängigen Kontrolle unterworfen. Dies kommt durch die unabhängige Bewertung der Konformitätsbewertungsstellen in Wahrnehmung ihrer Akkreditierung zustande (vgl. [9, 10 und 11]).

- **Risk treatment:** Various approaches are conceivable while trying to avoid the causes of errors. Simple and transparent models with a reduced number of parameters can be used in design and development with the goal of improving the transparency and explicability of AI systems. This is subjected to public control with the aim of securing the quality assurance chain. This is expressed by the independent assessment of conformity assessment bodies through their accreditation procedures (cf. [9, 10 and 11]).

2.8 Operating the AI system

In this phase, the AI system begins to be used in accordance with the defined geographical, behavioural or functional conditions ("intended use"). This corresponds to phases 11 (operation, maintenance and performance monitoring) in the development of railway applications. This phase is considered in its individual parts:

- **Potential failures and their impacts:** The failures in this phase are based on the fact that there can be a change in environmental parameters and input variables over time during operations. These changed operating conditions lead to a discrepancy between the input data and the data covered and taught during the training. Another possible failure may result from any possible changes in the AI functions during operations (model drift). In all cases, the AI system makes decisions that are unexpected by the developer.

Steuern, stellen, sichern.

Scheidt & Bachmann – innovative Sicherheitstechnologie seit 1872.

- Betriebsleittechnik
- Stellwerkstechnik
- Bahnübergangstechnik

Systeme für Signaltechnik

www.scheidt-bachmann.de

SCHEIDT&BACHMANN

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

2.8 Verwendung des KI-Systems

In dieser Phase beginnt die Verwendung des KI-Systems gemäß den festgelegten geografischen, verhaltensbezogenen oder funktionalen Rahmenbedingungen („bestimmungsgemäßer Gebrauch"). Dies entspricht in der Entwicklung von Bahnanwendungen der Phase 11 (Betrieb, Instandhaltung und Leistungsüberwachung). Diese Phase wird in ihren einzelnen Bestandteilen betrachtet:

- **Potenzielle Fehler und Fehlerauswirkung:** Fehler in dieser Phase beruhen darauf, dass es im Betrieb zu einer zeitlichen Veränderung von Umgebungsparametern und Eingangsgrößen kommen kann. Diese geänderten Betriebsbedingungen führen zu einer Diskrepanz von Eingabedaten zu den im Training angewendeten und eingelernten Daten. Ein weiterer möglicher Fehler resultiert aus ggf. veränderlichen KI-Funktionen im Betrieb (Model Drift). In allen Fällen trifft das KI-System nicht vom Entwickler erwartete Entscheidungen.
- **Fehlerursachen:** Abweichungen von Eingangs- und Trainingsdaten haben zwei Ursachen. Erstens kann es in der Phase des Betriebs zu unberechtigten Zugriffen Dritter kommen. Diese Zugriffe können sich zum einen auf die gezielte Manipulation von Eingabedaten beziehen (adversarial attack). Zum anderen kann ein Zugriff auf die Struktur und die Parameter des Modells intendiert sein (model stealing). Die bekannte Struktur des KI-Modells kann in der Folge dazu dienen, zukünftige Manipulationen von Eingabedaten noch besser an mögliche Schwachstellen des KI-Systems anzupassen. Zweitens kann es im Betrieb zu einer Veränderung von Umgebungsparametern und Eingangsgrößen im Zeitverlauf kommen (englisch: concept drift). Diese geänderten Bedingungen führen zu einer Diskrepanz von Eingabedaten zu den im Training abgedeckten und eingelernten Daten. Veränderliche Reaktionen des KI-Systems können aus dem eigenständigen Lernen des KI-Systems während des Betriebs resultieren. Die neuen (veränderten) Funktionen des KI-Systems bringen neue Risiken hervor, die bei Inverkehrbringen nicht vorgelegen haben. Dementsprechend ist auch der ursprüngliche Nachweis der zugesicherten Eigenschaften des KI-Systems nicht mehr gültig (vgl. Abschnitt 2.7).
- **Risikobehandlung:** Unberechtigte Zugriffe Dritter sind durch umfassende Schutzkonzepte zu unterbinden [8]. So können im Rahmen ganzheitlicher IT/OT-Security-Konzepte beispielsweise Anfragen an und Zugriffe auf das KI-System protokolliert werden und Protokolldaten regelmäßig auf Anomalien untersucht werden. Weiterhin müssen Prozesse etabliert werden, um auf Sicherheitsvorfälle im Betrieb zeitnah reagieren zu können. Hilfreich sind auch regelmäßige durchgeführte eigene Angriffe auf das KI-System (Red Teaming). Das Red Team ist eine unabhängige Gruppe, die als Gegner auftritt und zum Ziel hat, Sicherheitslücken vor einem externen Dritten aufzuspüren. Um eine Diskrepanz zwischen Trainings- und Eingangsdaten zu offenbaren, sollte die korrekte Funktionsweise des KI-Systems in regelmäßigen Abständen überprüft werden. Eine Veränderung von KI-Funktionen nach Inbetriebnahme wird in der Regel unterbunden. Andernfalls ist eine regelmäßige Kontrolle des KI-Systems im Betrieb durch den Menschen während des gesamte Lebenszyklus zu gewährleisten.

2.9 Beobachten des KI-Systems nach Inverkehrbringen

Diese Phase umfasst alle Tätigkeiten, die Anbieter von KI-Systemen zur proaktiven Sammlung und Überprüfung von Erfahrungen mit der Nutzung der von ihnen in Verkehr gebrachten KI-Systeme durchführen. Hierdurch soll festgestellt werden, ob Korrektur- oder Präventivmaßnahmen unverzüglich zu ergreifen sind. Für KI-Systeme kann diese Produktbeobachtung dadurch erfolgen, dass Bewertungsergebnisse zusammen mit den dazugehörigen Quelldaten (z. B. Sensordaten) gespeichert, an den Hersteller übertragen und dort ausgewer-

- **The causes of these failures:** The causes of any deviation between the input and training data are twofold. First, unauthorised access by third parties may occur during the operations phase. On the one hand, this access can refer to the targeted manipulation of input data (an adversarial attack). Access to the model's structure and parameters may also be intended (model stealing). The known structure of the AI model can subsequently be used to better adapt any future manipulation of the input data to possible weaknesses in the AI system. Secondly, environmental parameters and input variables can change over time during operations (concept drift). These changed operating conditions lead to a discrepancy between the input data and the data covered and learned during training. Changed reactions in the AI system can result from the AI system's independent learning during operations. The new (changed) functions of the AI system introduce new risks that did not exist when the system was placed on the market. Accordingly, the original evidence of the warranted characteristics of the AI system is no longer valid (cf. section 2.7).
- **Risk treatment:** Unauthorised access by third parties must be prevented by means of comprehensive protection concepts [8]. For example, requests and access to the AI system can be logged and the log data can be regularly examined for any anomalies. Furthermore, processes must be established to enable prompt reactions to security incidents during operations. It is also helpful to regularly carry out your own attacks on the AI system (red teaming). The red team is an independent group that acts as an adversary and aims to detect any security vulnerabilities before an external third party. The correct functioning of the AI system should be checked at regular intervals in order to reveal any discrepancies between the training and input data. The modification of AI functions after commissioning is usually prevented. Otherwise, regular human control of the AI system in operation should be implemented throughout its lifecycle.

2.9 Market surveillance of the AI system

This phase includes all the activities carried out by the AI system manufacturers to proactively collect and review the user experience pertaining to the AI systems they have placed on the market. This is to determine whether any corrective or preventive action needs to be taken immediately. For AI systems, this market surveillance can be done by storing evaluation results together with the associated source data (e.g. sensor data), transferring it to the manufacturer and evaluating it there. This phase should also be assigned to phase 11 for railway applications:

- **Potential failure and failure effect:** Malfunctions may not be detected in this phase. If any malfunctions are detected, the causes behind them may not be explained. Furthermore, corrective actions by the AI system manufacturers may not be implemented or may be inadequate.
- **The causes of these failures:** On the one hand, the causes can be found in the insufficient implementation of market surveillance or the obligation for effective hazard control on the part of the manufacturer. On the other hand, the principle-related lack of transparency and explicability of AI models also has a negative effect here. There may be uncertainties with regard to the required period for timely rectification during the implementation of corrective measures.
- **Risk treatment:** The description of a plan for the implementation of market surveillance has already been subjected to

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

tet werden. Auch diese Phase ist der Phase 11 für Bahnanwendungen zuzuordnen:

- **Potenzielle Fehler und Fehlerauswirkung:** In dieser Phase werden gegebenenfalls Fehlfunktionen nicht erkannt. Sind Fehlfunktionen erkannt worden, können die hinter diesen Fehlfunktionen liegenden Ursachen möglicherweise nicht erklärt werden. Des Weiteren werden möglicherweise keine oder nur unzureichende Korrekturmaßnahmen von den Herstellern der KI-Systeme umgesetzt.
- **Fehlerursachen:** Ursachen liegen zum einen in einer unzureichenden Umsetzung der Marktbeobachtung bzw. sind in der Verletzung der Pflicht zur effektiven Gefahrsteuerung seitens des Herstellers begründet. Zu anderen wirken sich auch hier die prinzipbedingte mangelnde Transparenz und vollumfängliche Erklärbarkeit von KI-Modellen negativ aus. Für die Umsetzung von Korrekturmaßnahmen bestehen ggf. Unsicherheiten in Bezug auf die erforderliche Frist zur zeitgerechten Behebung.
- **Risikobehandlung:** Die Beschreibung eines Plans zur Durchführung der Marktbeobachtung wird bereits vor der Inbetriebnahme einer unabhängigen Konformitätsbewertung unterworfen (vgl. Abschnitt 2.2). Außerdem wirken auch hier die zuvor beschriebenen Ansätze der Komplexitätsreduktion von KI-Modellen positiv (vgl. Abschnitt 2.7). Darüber hinaus muss sich die Konformitätsbewertung über den gesamten Lebenszyklus erstrecken und somit auch den Zeitraum nach Inverkehrbringen des KI-Systems umfassen. Für die zeitgerechte Umsetzung von Korrekturmaßnahmen geben in der Bahnindustrie etablierte Vorgehensweisen zur risikoorientierten Beurteilung potenzieller Sicherheitsmängel eine Orientierung [12].

2.10 Wartung und Pflege des KI-Systems

Diese Phase fokussiert die Gewährleistung des ordnungsgemäßen Betriebs dieses KI-Systems, auch in Bezug auf Software-Updates. Auch diese Phase ist der Phase 11 für Bahnanwendungen zuzuordnen:

- **Potenzielle Fehler und Fehlerauswirkung:** Mögliche Fehler resultieren in dieser Phase aus einer nicht sachgemäßen Umsetzung wesentlicher Änderungen von KI-Systemen durch den Hersteller. Weitere Fehler resultieren aus der unzureichenden Wartung und Pflege des KI-Systems auch in Bezug auf Software-Updates. Beide Ursachen stellen Abweichungen vom bestimmungsgemäßen Betrieb des KI-Systems dar. Die Aussagen der zum Zeitpunkt des Inverkehrbringens vorliegenden Konformitätsbewertung gelten nicht mehr vollinhaltlich. Ein vertrauenswürdiger Betrieb des KI-Systems kann daher nicht mehr angenommen werden.
- **Fehlerursachen:** Abweichungen in der Umsetzung wesentlicher Änderungen liegen möglicherweise in der Unkenntnis des systematischen Prozesses zur Behandlung wesentlicher Änderungen begründet. Unzulänglichkeiten in der Wartung und Pflege des KI-Systems liegen möglicherweise in einer unvollständigen Nutzerdokumentation begründet oder in der mangelhaften Umsetzung der sicherheitsbezogenen Anwendungsregeln durch den Betreiber des KI-Systems.
- **Risikobehandlung:** Für wesentliche Änderungen bestehen für Bahnanwendungen etablierte Vorgehensweisen, die auch eine unabhängige Bewertung des Risikomanagementverfahrens umfassen [13]. Die vollständige und korrekte Weiterleitung der sicherheitsbezogenen Anwendungsregeln an die Wartung und Instandhaltung ist Bestandteil der Nutzerdokumentation, welche im Rahmen des Konformitätsbewertungsverfahrens zur Inbetriebnahme geprüft wird.

2.11 Außerbetriebsetzung des KI-Systems

Diese Phase des Lebenszyklus beschreibt die Abkündigung des KI-Systems sowie den Übergang zu einem neuen KI-System. In dieser Phase

TrackMedia

Infrastrukturprojekte 2022

Mit dem neu erschienenen „Infrastrukturprojekte 2022 – Bauen für die starke Schiene“ **gibt die Deutsche Bahn zum siebten Mal einen Einblick in die Investitionen** zur Stärkung der Schiene.

NEU

1. Auflage Okt. 2022,
Hrsg. DB Netz AG,
198 Seiten,
ISBN 978-3-96245-253-7,
Print € 49,-*
www.trackomedia.com/infra22

Mehr Infos und Bestellung: www.trackomedia.com

Handbuch Das System Bahn
Print mit E-Book Inside € 99,-*
www.trackomedia.com/systembahn

ETCS in Deutschland
Print mit E-Book Inside € 79,-*
www.trackomedia.com/etcsdeutschland

* Preise inkl. MwSt, zzgl. Versand

BESTELLUNGEN:
Tel.: +49 7953 718-9092
Fax: +49 40 228679-503
E-Mail: office@trackomedia.com
Online: www.trackomedia.com

PER POST:
GRT Global Rail Academy and Media GmbH | Trackomedia
Kundenservice
D-74590 Blaufelden

Unsere Bücher erhalten Sie auch im gut sortierten Buchhandel.

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

kann ein KI-System gelöscht oder signifikant verändert werden, sodass ein neues KI-System erstellt wird. Damit beginnt ein neuer Lebenszyklus. Dies entspricht in der Entwicklung von Bahnanwendungen bei konventionellen Systemen der Phase 12 (Außerbetriebnahme).

3 Fazit und Ausblick

Die Einführung von KI-Systemen für sicherheitsrelevante Bahnanwendungen ist eine disruptive Innovation. Gemäß einer Verbraucherstudie des Verbandes der Technischen Überwachungsvereine (VDTÜV) besteht in der Bevölkerung eine grundsätzlich positive Grundeinstellung zu KI [14]. Allerdings stellt die Studie auch Sorgen und Ängste fest. Wenn KI in immer mehr Lebensbereiche vordringt, muss Vertrauen geschaffen werden:

- **Risikoakzeptanz:** Je höher das Risiko einer Anwendung, desto strenger müssen die Regeln hierfür sein. 58 % der Deutschen begrüßen den risikobasierten Ansatz der Kommission der Europäischen Union [6].
- **Stärkere Unabhängigkeit in der Prüfung:** 66 % der Befragten haben eher großes bis sehr großes Vertrauen in unabhängige akkreditierte Sachverständigenorganisationen. Damit liegt das Vertrauen in diese Organisationen klar vor den Herstellern (55 %) oder staatlichen Stellen (49 %).
- **Stärkere Rolle der Produktbeobachtung:** 81 % der Befragten sprechen sich dafür aus, dass die Sicherheit von Produkten und Anwendungen mit KI vor der Markteinführung von herstellerunabhängigen Stellen geprüft werden sollte. 79 % sind der Meinung, dass solche Prüfungen auch nach Inverkehrbringen erforderlich sind [14].
- **Erfordernis harmonisierter Normen:** Produkt- und Prozessprüfungen müssen auf einheitlichen Standards beruhen. Mit der „Normungsroadmap Künstliche Intelligenz" liegt eine umfassende Analyse des Bestands und des Bedarfs an internationalen Normen und Standards für KI vor [15]. ■

an independent conformity assessment prior to commissioning (cf. section 2.2). In addition, the previously described approaches to the complexity reduction of AI models should also have a positive effect here (cf. section 2.7). Furthermore, the conformity assessment should cover the entire lifecycle and thus also the time after the AI system has been placed on the market. Established procedures in the railway industry for the risk-oriented assessment of potential safety deficiencies provide an orientation for the timely implementation of corrective measures [12].

2.10 Maintaining the AI system

This phase focuses on ensuring the proper functioning of this AI system, also in terms of software updates. This phase is also assigned to phase 11 for railway applications:

- **Potential failures and their impacts:** Potential failures in this phase result from the improper implementation of any significant changes to the AI systems by the manufacturer. Further failures result from inadequate maintenance of the AI system, including software updates. Both causes represent deviations from the intended operation of the AI system. The conformity assessment statements available at the time of placement on the market are no longer valid. The trustworthy operation of the AI system can no longer be assumed.
- **The causes of these failures:** Deviations in the implementation of any significant changes may be due to ignorance of the systematic process for handling significant changes. Inadequacies in the maintenance and care of the AI system may be due to incomplete user documentation or poor implementation of the safety-related application rules by the AI system operator.
- **Risk treatment:** Established procedures exist for significant changes to railway applications, including an independent assessment of the risk management process [13]. The complete and correct forwarding of the safety-related application conditions to maintenance and repair is part of the user documentation, which is checked as part of the conformity assessment procedure prior to commissioning.

2.11 Decommissioning the AI system

This phase of the lifecycle describes the discontinuation of the AI system and the transition to a new AI system. In this phase, an AI system can be deleted or significantly changed so that a new AI system is created. This marks the beginning of a new lifecycle. It corresponds to phase 12 (decommissioning) in the development of railway applications

3 Conclusion and outlook

The introduction of AI systems for safety-relevant railway applications is a disruptive innovation. According to a consumer study by the Verband der Technischen Überwachungsvereine (VDTÜV), there is a fundamentally positive basic attitude towards AI among the population [14]. However, the study also notes concerns and fears. If AI penetrates into more and more areas of life, trust must be established:

- **Risk acceptance:** The higher the risk of an application, the stricter the rules for it must be. 58 % of Germans welcome the European Union Commission's risk-based approach [6].
- **Greater independence in assessment:** 66 % of the respondents have somewhat big to very big trust in independent accredited expert organisations. This places the trust in these

Homepageveröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH

LITERATUR | LITERATURE
[1] Kommission der Europäischen Union: Weissbuch „Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen", COM(2020) 65 final, 19.02.2020
[2] Schnieder, E.: Methoden der Automatisierung – Beschreibungsmittel, Modellkonzepte und Werkzeuge für Automatisierungssysteme
[3] DIN SPEC 92001-1 Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model; April 2019
[4] DIN EN 50126-1:2018-10: Bahnanwendungen – Spezifikation und Nachweis von Zuverlässigkeit, Verfügbarkeit, Instandhaltbarkeit und Sicherheit (RAMS) – Teil 1: Generischer RAMS-Prozess, Deutsche Fassung EN 50126-1:2017
[5] DIN EN 60812 (Entwurf): Fehlzustandsart- und -auswirkungsanalyse (FMEA) (IEC 56/1579/CD:2014)
[6] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, 2021/0106 (COD)
[7] ISO/PAS 21448: Road vehicles – Safety of the intended functionality
[8] DIN EN ISO/IEC 27001:2017-06: Informationstechnik – Sicherheitsverfahren – Informationssicherheitsmanagementsysteme – Anforderungen (ISO/IEC 27001:2013 einschließlich Cor 1:2014 und Cor 2:2015), Deutsche Fassung EN ISO/IEC 27001:2017
[9] Röhl, H.-C.: Akkreditierung und Zertifizierung im Produktsicherheitsrecht, Springer (Berlin) 2000
[10] Ernsthaler, J.; Strübbe, K.; Bock, L.: Zertifizierung und Akkreditierung technischer Produkte – Ein Handlungsleitfaden für Unternehmen, Springer Verlag (Berlin) 2007
[11] Schnieder, L.: Öffentliche Kontrolle der Qualitätssicherungskette für einen sicheren und interoperablen Schienenverkehr, in: Eisenbahntechnische Rundschau, 4/2017, S. 38 – 41
[12] DIN VDE V 0831-100:2019-08: Elektrische Bahn-Signalanlagen Teil 100: Risikoorientierte Beurteilung von potenziellen Sicherheitsmängeln und risikoreduzierende Maßnahmen
[13] Durchführungsverordnung (EU) Nr. 402/2013 der Kommission vom 30. April 2013 über die gemeinsame Sicherheitsmethode für die Evaluierung und Bewertung von Risiken und zur Aufhebung der Verordnung (EG) Nr. 352/2009
[14] TÜV-Verband e.V: Sicherheit und Künstliche Intelligenz – Erwartungen, Hoffnungen, Risiken. Repräsentative Befragung der Bevölkerung in Deutschland im Auftrag des TÜV-Verbands (August 2021)
[15] Deutsches Institut für Normung: Deutsche Normungsroadmap Künstliche Intelligenz. Berlin, November 2020

organisations clearly ahead of the manufacturers (55 %) or government agencies (49 %).

- **A stronger role for market surveillance:** 81 % of respondents are in favour of the safety of products and applications with AI being tested by manufacturer-independent bodies before they are placed on the market. 79 % are of the opinion that such tests are also required after the products have been placed on the market [14].
- **The need for harmonised standards:** Product and process testing must be based on uniform standards. The Artificial Intelligence Standardisation Roadmap provides a comprehensive analysis of the stock and need for international norms and standards for AI [15]. ■

AUTOR | AUTHOR

PD Dr.-Ing. habil. Lars Schnieder
Geschäftsführer / *Chief Executive Director*
ESE Engineering und Software-Entwicklung GmbH
Anschrift / *Address*: Am Alten Bahnhof 16, D-38122 Braunschweig
E-Mail: lars.schnieder@ese.de

Homepage veröffentlichung unbefristet genehmigt für ESE Engineering und Software-Entwicklung GmbH / Rechte für einzelne Downloads und Ausdrucke für Besucher der Seiten genehmigt / © DVV Media Group GmbH