Anwendung Künstlicher Intelligenz in der Bahntechnik für sicherheitsrelevante Anwendungen – Chancen und Probleme

The application of artificial intelligence in railway technology for safety-relevant applications – opportunities and problems

Jens Braband | Hendrik Schäbe

Die Erfolge der Anwendungen von sogenannter „Künstlicher Intelligenz" (KI) sind unübersehbar. KI wird auf zahlreichen Gebieten eingesetzt: Unterhaltung, Kunst und in vielen technischen Systemen sind KI-gestützte Assistenzsysteme im Einsatz. Einen guten Überblick über KI-Methoden und -Einsatzfelder findet man z. B. in [1]. Und auch für den Einsatz von KI in sicherheitsrelevanten Systemen werden aufwendige technische und politische Vorbereitungen getroffen, bis hin zu einem geplanten „TÜV für KI". Trotz aller Erfolge ist es allerdings bisher nicht gelungen, einen Sicherheitsnachweis für eine KI zu führen oder gar eine Zulassung zu erwirken, wenn die KI allein eine sicherheitsrelevante Entscheidung treffen soll. Sowohl in der Medizin- als auch in der Automobiltechnik muss die Entscheidung der KI außerhalb von Testfeldern immer von einem Menschen überstimmt werden können. Und gerade in der Automobiltechnik gab es trotzdem bereits tödliche Unfälle, da entweder die KI nicht bestimmungsgemäß eingesetzt wurde oder der Sicherheitsfahrer abgelenkt war. Oder es gibt bisher unerklärte Vorfälle, wie die Kollisionen von autonom fahrenden Autos mit stehenden Einsatzfahrzeugen mit Blinklicht. Es ist daher eine gewisse Ernüchterung eingetreten, und zumindest die deutschen Automobilhersteller gehen in diesem Jahrzehnt nicht mehr davon aus, dass es fahrerlose, universell einsetzbare Autos geben wird.

1 Einleitung

Auch in der Eisenbahntechnik ist die Diskussion über KI-gestützte Systeme in vollem Gange, insbesondere beim automatisierten Fahren. Gegenüber anderen Anwendungsfeldern besitzt das Eisenbahnsystem den Vorteil, dass fahrerlose Systeme bei U-Bahnen bereits den Normalfall darstellen und dass die Einsatzbedingungen einfacher sind als z. B. in der Automobiltechnik.
Weitere mögliche Anwendungen kann man bei der Bahn z. B. in den folgenden Bereichen sehen [2]:

- Fahrerassistenzsysteme wie Warnanlagen,
- energetisch optimierte Routenführung und
- prädiktive Wartung.

Dieser Beitrag beleuchtet die grundsätzlichen Probleme, die gelöst werden müssen, bevor man einer KI die alleinige Sicherheitsverantwortung für eine Entscheidung überlassen darf. Die Darstellung soll

The success of applications of so-called "artificial intelligence" (AI) is unmistakable. AI is used in numerous fields: entertainment and art, while AI-supported assistance systems are also in use in many technical systems. A good overview of AI methods and fields of application can be found in [1]. In addition, technical and political preparations are also being prepared for the use of AI in safety-relevant systems all the way up to a planned "TÜV for AI". Despite all the success, however, it has not yet been possible to provide an AI safety case or even to obtain approval for cases where AI is solely responsible for safety-relevant decisions. If used outside a test field, in both medical and automotive technology, an AI decision must always be able to be overruled by a human operator. This is especially the case in automotive technology. Despite this, however, there have already been fatal accidents, because either the AI was not used as intended or the safety driver was distracted. In addition, there have also been previously unexplained incidents, such as collisions of autonomous driving cars with stationary emergency vehicles showing flashing lights. As such, a certain degree of disillusionment has occurred. At the very least, German car manufacturers no longer assume that driverless, universally usable cars will be available within this decade..

1 Introduction

Discussions about AI-supported systems have also been raised in railway technology, especially in automated driving. Compared to other fields of application, the railway system has the advantage that driverless systems are already the norm in metros and that the conditions of use are simpler than e.g. in automotive technology.
Other possible applications can be seen in railways, e.g., in the following areas [2]:

- driver assistance systems, such as warning systems
- energy optimised routing and
- predictive maintenance

This article highlights the fundamental problems that must be resolved before AI can be made responsible for safety decisions. The presentation is intended to be simple. This imposes a number of constraints:

dabei einfach aufgebaut werden. Das bedingt eine Reihe von Einschränkungen:

- In diesem Beitrag beschränkt sich die Diskussion auf eine besonders beliebte und erfolgreiche Klasse von KI, das sogenannte Maschinelle Lernen (ML), und in manchen Aspekten auch nur auf dessen populärste Ausprägung, die Neuronalen Netzwerke (NN).
- Es wird davon ausgegangen, dass die KI eine sicherheitsrelevante Entscheidung selbstständig treffen muss, d. h. eine Klassifikationsaufgabe ohne menschliche Unterstützung. D. h. die Diskussion gilt nicht für Assistenzsysteme o. ä.
- Es werden keine selbstlernenden KI-Systeme betrachtet, die sich im Betrieb dynamisch weiter verändern, sondern nur solche, die vor Beginn des Einsatzes trainiert werden und dann unverändert eingesetzt werden. Bei selbstlernenden Systemen ist die Problematik weitaus komplizierter, und auch die hier herangezogene Aufgabenstellung passt dafür nicht.

Zu guter Letzt muss man auch noch anmerken, dass es sich nur um eine möglichst allgemeinverständliche Annäherung an ein komplexes Problem handelt und dabei gewisse Vereinfachungen in Kauf genommen werden müssen. KI-Experten werden daher um Nachsicht gebeten, wenn einige technische Details nur verkürzt dargestellt werden können. In einer anderen Publikation wird etwas stärker auf technische Details eingegangen [3].

2 Was ist Künstliche Intelligenz?

Es gibt viele Veröffentlichungen, und viele Systeme werden als künstlich intelligent bezeichnet. Eine Übersicht findet sich z. B. bei Brunette et al. [4]. Ausgangspunkt war der Turing-Test in den 1950er Jahren, mit dem überprüft werden sollte, ob ein Computer ein intelligentes Verhalten zeigt, das mit dem eines Menschen vergleichbar ist. Später wurde das Konzept der evolutionären Programme eingeführt. Der Begriff „Künstliche Intelligenz" wurde erstmals 1956 am Dartmouth College verwendet. In der Zwischenzeit wurden von vielen Forschern verschiedene Konzepte vorgeschlagen.

Aber trotzdem hat sich in der Normung weder national noch international eine einheitliche Definition durchgesetzt. Auch die Normungs-Roadmap „Künstliche Intelligenz" des DIN und der DKE [5] vermeidet eine explizite Definition, benennt aber konkrete Handlungsbedarfe, z. B. „Horizontale KI-Basis-Sicherheitsnorm erstellen". Maschinelles Lernen wird dort als *„Technik, die ein System in die Lage versetzt, aus Daten und Interaktionen zu lernen*" definiert.

KI kann als Intelligenz definiert werden, die von Maschinen gezeigt wird. KI ahmt kognitive Funktionen, Lernen, Problemlösung usw. nach.

Es gibt bisher kein veröffentlichtes vollständiges Sicherheitsargument für KI-Anwendungen, aber es gibt viele Forschungsprojekte zu Sicherheitsbegründungen für KI.

In jüngster Zeit wurden jedoch einige Ansätze aus sicherheitstechnischer Sicht vorgeschlagen, vor allem der Entwurf der Norm UL 4600 [6], der einen Sicherheitsansatz für autonome Fahrzeuge fordert, die KI-Algorithmen nutzen können. Allerdings wird auch in der UL 4600 nur das „What to argue", nicht aber das „How" behandelt.

Andere Normungsgremien, z. B. die deutsche DKE, setzen auf einen prozess- und lebenszyklusorientierten Ansatz. Putzer [7] propagierte ein sogenanntes λAI, ein Maß ähnlich einer Gefährdungsrate in der funktionalen Sicherheit, gibt aber keine präzise Definition.

In Zusammenhang mit der zunehmenden Verbreitung von KI-Systemen hat der Begriff der „erklärbaren KI" (explainable artificial intelligence) eine breite Verwendung gefunden [8]. Dies bedeutet, dass man nachvollziehen möchte, wie eine KI entschieden hat. Letztendlich bedeutet dies, dass man sich das zugrunde liegende mathema-

- The discussion in this article is limited to one particularly popular and successful class of AI called machine learning (ML) and in some aspects only to its most popular manifestation, i.e. neural networks (NN).
- It is assumed that the AI must be able to make a safety-relevant decision independently, i.e. a classification task without human support. As such, the discussion does not apply to assistance systems or the like.
- No self-learning AI systems that continue to change dynamically during operations have been considered. We have only considered systems that are trained before being commissioned and are then are left unchanged. The problem is far more complicated in the case of self-learning systems and the task used here is likewise unsuitable for this.

Last but not least, it must also be noted that we have only endeavoured to give a general understanding of a complex problem in this paper and that certain simplifications must be accepted in the process. AI experts are therefore asked to bear with us, if some of the technical details have only been presented in abbreviated form. The technical specifics are dealt with in more detail in another publication [3].

2 What is artificial intelligence?

There are many publications and many systems are designated as being artificially intelligent. An overview can be found in Brunette et al. [4]. The starting point was the Turing test in the 1950s, which was intended to test whether a computer exhibited intelligent behaviour comparable to that of a human. Later, the concept of evolutionary programs was introduced. The term "artificial intelligence" was first used at Dartmouth College in 1956. In the meantime, various concepts have been proposed by many researchers.

Nevertheless, no uniform definition has been established in standardisation, either nationally or internationally. The "Artificial Intelligence" standardisation roadmap of DIN and DKE [5] also avoids an explicit definition, but specifies concrete needs for action, e.g. "create a horizontal basic AI safety standard". Machine learning is defined there as "*technology that enables a system to learn from data and interactions*".

AI can be defined as intelligence displayed by machines. AI mimics cognitive functions, learning, problem solving, etc.

No complete safety argument has been published for an AI application, but there have been many research projects on the safety justifications for AI.

However, some approaches have recently been proposed from a safety point of view, most notably the draft UL 4600 standard [6], which calls for a safety approach to autonomous vehicles that can use AI algorithms. However, UL 4600 also only elaborates "What" to argue, but not the "How".

Other standardisation committees, e.g. the German DKE, have focussed on a process and lifecycle-oriented approach. Putzer [7] has propagated a so-called λAI, a measure similar to a hazard rate in functional safety, but has not given a precise definition of it.

The term "explainable AI" (explainable artificial intelligence) has found widespread use in connection with the increasing proliferation of AI systems; see [8]. This ultimately means that one wants to understand how the AI has reached its decision. Likewise, it means that it is necessary to look at the underlying mathematical model. In many cases, machine learning can be traced back to statistical models; see [9]. The mathematical foundations in question, summarised as so-called statistical learning theory, are even relatively old [10]. The peculiarity of these models, however, is that the op-

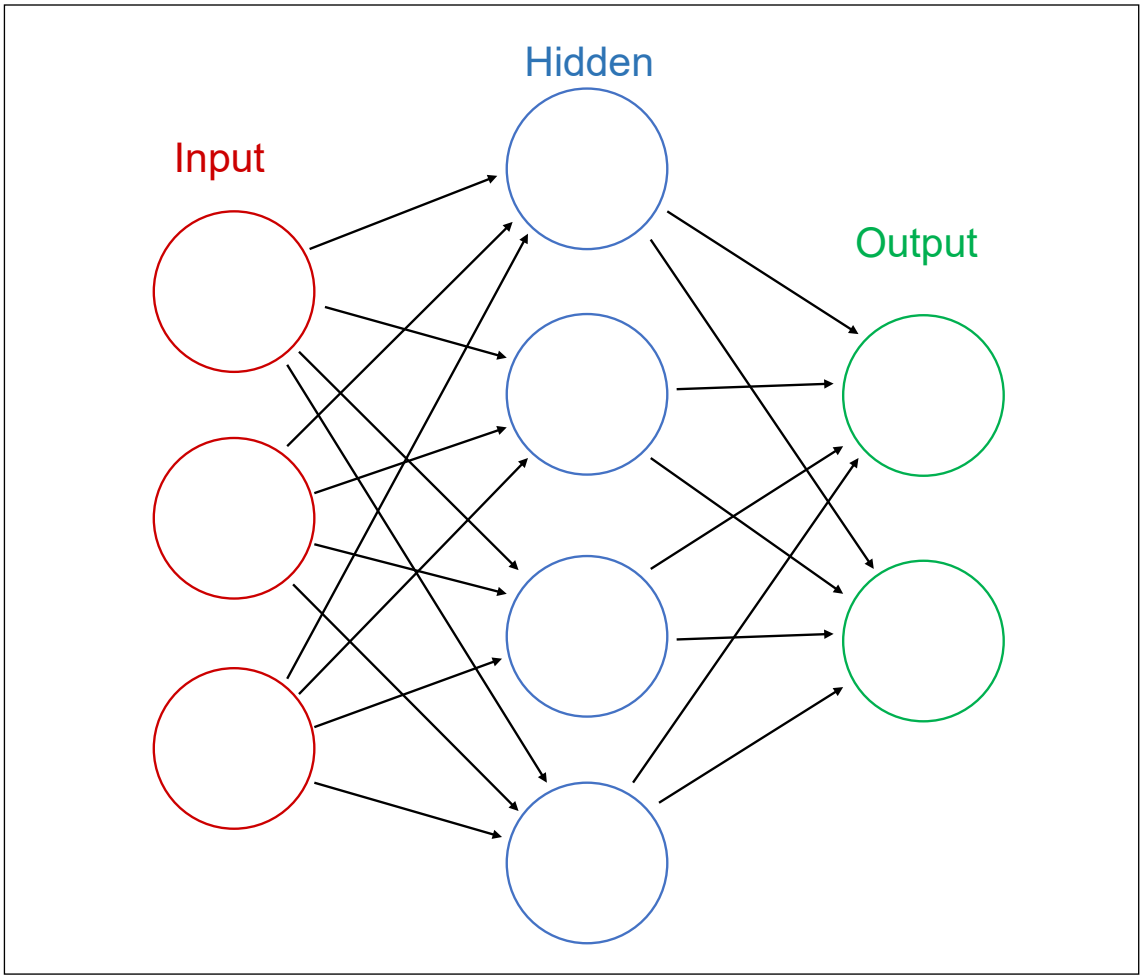

Bild 1: Künstliches NN mit einer versteckten Schicht
Fig. 1: An artificial NN with a hidden layer
Quelle / *Source*: User: Glosser.ca / Wikimedia Commons / CC-BY-SA-3.0

tische Modell ansehen möchte. In vielen Fällen lässt sich Maschinelles Lernen auf statistische Modelle zurückführen [9]. Die betreffenden mathematischen Grundlagen, zusammengefasst als sog. Statistische Lerntheorie, sind sogar schon relativ alt [10]. Die Besonderheit an diesen Modellen ist jedoch, dass die Operationen in sehr hochdimensionalen Räumen durchgeführt werden, sodass das Modell eine Vielzahl innerer Parameter besitzt. Dadurch wird aber auch deutlich, dass KI nicht denkt – sondern einfach Daten analysiert. Zusätzliche Informationen, die ein Mensch hat, kann die KI nicht verwenden. Dies legt nahe, statt der modernen Begriffe wieder die klassische Terminologie wie z. B. Mustererkennung zu verwenden, denn häufig laufen die Anwendungen genau darauf hinaus.

3 Eine Analogie für KI-Anwendungen

Die derzeit beliebteste und erfolgreichste Anwendung des ML sind die sogenannten künstlichen Neuronalen Netzwerke (NN). NN versuchen, die Struktur des menschlichen Gehirns nachzubilden, indem einfachste Nervenzellen (sog. Perzeptrons) modelliert und verknüpft werden. In der grafischen Darstellung der allereinfachsten Ausprägung (vgl. Bild 1) besteht ein NN aus einer Input-Schicht, die die Eingaben aufnimmt, einer versteckten („hidden“) Schicht, in der für den Benutzer nicht sichtbare Berechnungen erfolgen, sowie eine Output-Schicht, die dann wieder sichtbare Ergebnisse darstellt.
Kompliziertere NN (oft „tiefe“ NN genannt) unterscheiden sich in der Anzahl der versteckten Schichten sowie der Anzahl N der

erations are performed in very high-dimensional spaces, so that the model has a large number of internal parameters. However, this also makes it clear that AI does not think - it simply analyses data. AI cannot use the additional information that a human has. This suggests that classic terminology such as pattern recognition should be used again instead of the modern terms, because this is often exactly what the applications boil down to.

3 An analogy for AI applications

Currently, the most popular and successful applications of ML are so-called artificial neural networks (NN). NN attempt to emulate the structure of the human brain by modelling and linking the simplest nerve cells (so-called perceptrons). In the graphic representation of the simplest form (cf. fig. 1), an NN consists of an input layer that receives the input, a hidden layer where calculations take place that are not visible to the user and an output layer that then presents the visible results.
More complicated NN (often called “deep” NN) differ in their number of hidden layers as well as the number of N nodes in the layers. The theory is described in [24].
The ML with NN presents itself statistically as an estimation problem for an unknown function [3].

4 Defining a SIL for functions with AI

In this section we will discuss, whether we need a Safety Integrity Level (SIL) for AI and if so, how it should be determined.

Knoten in den Schichten. Die theoretischen Grundlagen hierzu sind in [24] beschrieben.
Das ML mit NN stellt sich statistisch als ein Schätzproblem für eine unbekannte Funktion dar [3].

4 Definition eines SIL für Funktionen mit AI

In diesem Abschnitt werden wir erörtern, ob wir ein Sicherheitsintegritätsniveau für KI benötigen und wenn ja, wie es bestimmt werden sollte.
Das Konzept der Sicherheitsintegritätsstufe (SIL, Safety Integrity Level) wird in vielen Normen für funktionale Sicherheit verwendet. Die Mutternorm ist die bekannte IEC 61508. Für die Bestimmung von SIL sei auf Schäbe [13] verwiesen.
Bild 2 zeigt die Situation bei einem elektrischen, elektronischen, programmierbaren elektronischen System (E/E/PE-System). Hier haben wir ein zu steuerndes Gerät, Informationen von Sensoren, die in das Steuerungssystem eindringen, und Aktoren, die vom Steuerungssystem bedient werden. Je nach den Folgen eines fehlerhaften Verhaltens des Steuerungssystems erhält dieses einen Sicherheitsintegritätsgrad (SIL). Es spielt keine Rolle, welche Art von Steuerungssystem wir haben, dies kann auch ein KI-System sein.
Für die Risikoanalyse und die Bestimmung des SIL wird es ohnehin als Black Box betrachtet.
Ein SIL kann nun erforderlich sein, wenn das KI-System sicherheitsrelevante Aufgaben erfüllt, und der SIL kann mit denselben Methoden wie für ein E/E/PE-System bestimmt werden. Lediglich die Regeln für die Bewertung des SIL können je nach Art des Systems, das die Blackbox implementiert, unterschiedlich sein.
Mit welchem SIL müssten wir bei verschiedenen KI-Anwendungen rechnen? Dies würde vor allem von den Fehlerfolgen abhängen und davon, ob andere Risikominderungen möglich sind:

- Datenverarbeitung – hängt von den Ergebnissen ab und davon, was mit ihnen gemacht wird
- Assistenzsysteme – in der Regel kein SIL, wenn ein Mensch das System jederzeit außer Kraft setzen kann
- Spracherkennung – hängt davon ab, was mit dem Ergebnis gemacht wird und ob es sichere Backups gibt
- Gesichtserkennung – abhängig davon, was mit dem Ergebnis gemacht wird, d. h. welche Funktionen aktiviert werden
- Pflegeroboter – Verabreichung von Medikamenten, Tragen von Patienten, daher wäre sicherlich ein SIL erforderlich
- Autonome Fahrsysteme – können zu Unfällen führen, daher wäre ein SIL erforderlich.

In jedem Fall muss eine Risikoanalyse durchgeführt werden, um den SIL zu bestimmen – oder die Tatsache, dass es nicht notwendig ist, einen zu bestimmen. Die entsprechende Norm zur funktionalen Sicherheit muss angewendet werden.

The concept of a SIL is used in many standards for functional safety. The mother standard is the well-known IEC 61508. The reader can be referred to Schäbe [13] for the determination of SILs.
Fig. 2 shows the situation with a conventional electric, electronic, programmable electronic system (E/E/PE system). Here, we have equipment under control, information from sensors that enter the control system and actors operated by the control system. This system receives a SIL depending on the consequences of any faulty behaviour in the control system. Now, it no longer matters what type of control system we have and it might as well be an AI system.
It is considered as a black box anyway for the hazard analysis and the SIL determination.
Therefore, a SIL can be necessary, if the AI system performs safety relevant tasks and the SIL can be determined using the same methods as for an E/E/PE system. Only the rules for the assessment of the SIL may differ depending on the type of system that implements the black box.
What SIL would we have to expect for different AI applications? This would mainly depend on the consequences of any failure and whether or not any other risk mitigations are possible:

- Data processing – it depends on the results and what is done with it
- Assistance systems – normally without an SIL, if a human can always override the system
- Speech recognition – it depends on what is done with the result and whether there are any safe backups
- Face recognition – it depends on what is done with the result, i.e. which functions are activated
- Nursing robots – they dispense medicine and carry patients, so surely an SIL would be required
- Autonomous driving systems – can lead to accidents, so an SIL would be required.

In any case, a risk analysis needs to be carried out in order to determine the SIL – or the fact that it is not necessary to determine one. The relevant functional safety standard has to be applied.

5 Creating a safety case

If a SIL has been determined for a function, this SIL must also be demonstrated, if the function is implemented with AI. There are two conceivable ways of proving the safety of AI systems. The first one is the analytical way, while the second one would be the statistical way. Both ways will be briefly discussed in the following section.
Strictly speaking, the safety case consists of two parts that must cover two aspects. First, there is the hardware that the AI runs on and the software where it is implemented. We therefore need to prove a SIL appropriate for the application and the required tolerable functional failure rate (TFFR) [14]. In the case of the SIL, it

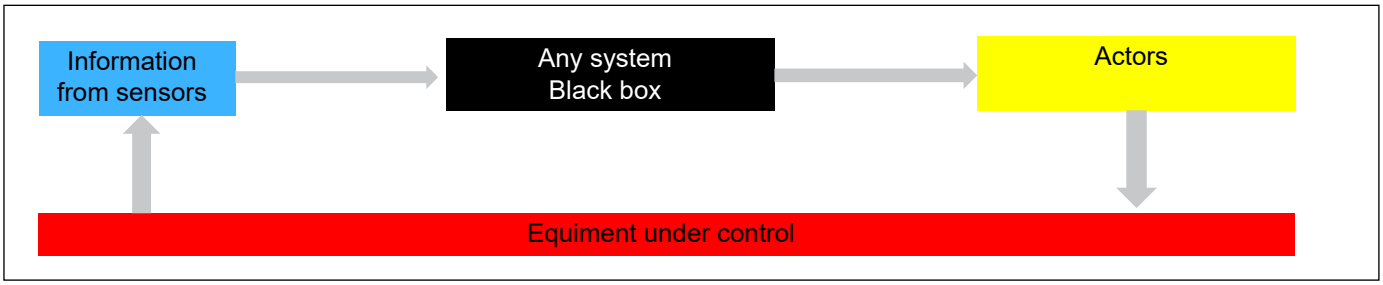

Bild 2: Beliebiges Kontrollsystem (schwarzer Kasten)
Fig. 2: An arbitrary control system (black box)

Quelle / *Source*: eigene Darstellung / *own illustration*

5 Erstellen eines Sicherheitsnachweises

Wenn für eine Funktion ein SIL ermittelt wurde, so ist dieser auch nachzuweisen, wenn die Funktion mit KI realisiert wird. Für den Nachweis der Sicherheit von KI-Systemen sind zwei Wege denkbar. Der erste ist der analytische Weg, der zweite wäre ein statistischer. Nachfolgend soll auf beide Wege kurz eingegangen werden.
Der Sicherheitsnachweis besteht dabei genau genommen aus zwei Teilen, die zwei Aspekte abdecken müssen. Zum einen gibt es eine Hardware, auf der die KI läuft und eine Software, in der sie implementiert ist. Wir müssten daher einen für die Anwendung entsprechenden SIL und die geforderte Gefährdungsrate (TFFR – Tolerable Functional Failure Rate) [14] nachweisen. Beim SIL ist zu beachten, dass ggf. die Tools, die zur Bestimmung der Parameter des NN verwendet wurden, auch sicherheitsrelevant sein könnten und dass die durch KI realisierte Funktion im Gesamtsystem validiert werden muss. Dabei wird in der Regel auch eine Sicherheitserprobung notwendig sein. Die Anforderungen des SIL bezüglich Software sollten leicht zu erfüllen sein, da im Prinzip nur eine relativ einfache Formel zu implementieren ist.
Zum anderen ist die KI selbst nicht deterministisch, sondern sie trifft Entscheidungen mit bestimmten Wahrscheinlichkeiten. Daher ist es sinnvoll, der KI einen Teil der TFFR zuzuordnen, d. h. der Rate des gefährlichen Versagens des technischen Systems, siehe [3] und [2].
Der erste Teil des Sicherheitsnachweises ist relativ einfach durchzuführen, schwieriger wird der zweite.

5.1 Analytischer Ansatz

Bei der Validierung der KI gibt es interessante Fragestellungen: Wenn z. B. bei der Validierung der KI ein Fehler auftritt, dann wird bei der Anwendung auch wieder ein Fehler auftreten, wenn derselbe Datensatz auftritt oder mit hoher Wahrscheinlichkeit ein Fehler, wenn ein „ähnlicher" Datensatz vorkommt. Dies könnte man als systematischen Fehler der KI-Funktion interpretieren. Auch wenn der Anteil von Fehlern sehr gering ist, darf man ein sicherheitsrelevantes System in Betrieb nehmen, von dem man weiß, dass es bei gewissen Datensätzen fehlerhaft reagiert? Oder muss man diese Datensätze sowie ihre „Umgebung" ausnehmen? Oder muss man gar fordern, dass die Validierung fehlerfrei verläuft? Dies könnte allerdings Probleme machen, wenn bei der Sicherheitserprobung oder im Betrieb Fehler auftreten. D. h. man wird vermutlich Fehler tolerieren müssen und damit eine gewisse Gefährdungsrate, die sich aufgrund von Fehlentscheidungen der KI-Funktion ergibt.
Bei dem Nachweis der TFFR ist zu berücksichtigen, dass die Forderungen in der Regel so streng sind, dass sie nur für geringe Sicherheitslevel, z. B. SIL1, durch praktischen Nachweis, z. B. Test oder Betriebserfahrung, nachgewiesen werden können. Für höhere SIL werden in der Regel analytische Nachweise nötig sein. Wenn man sich anschaut, wie solche Nachweise in anderen Bereichen erfolgen, z. B. für Rechnerarchitekturen oder Datenübertragung, dann erkennt man, dass solche Nachweise nur gelingen, wenn man ein sinnvolles Modell findet. Diese Erkenntnis ist nicht neu, auch der Turing-Preisträger Pearl [15] hat schon geschlussfolgert, dass es notwendig ist, "*to formulate a model of the process that generates the data, or at least some aspects of that process*". D. h. man muss zumindest teilweise verstehen, welche Struktur bzw. Eigenschaften die Daten haben. Erst dann kann man erwarten, dass man mittels analytischer Ergebnisse aus dem (Teil-)Modell Eigenschaften ableiten kann, die für höhere SIL die Anforderungen eines Sicherheitsnachweises erfüllen. Praktisch bedeutet dies, dass ein solcher Nachweis für Daten einfacher sein wird, die eine gewisse Struktur besitzen, z. B. aufgrund von physikalischen Eigenschaften der Datenquellen wie dies z. B. bei

should be noted that, if applicable, the tools used to determine the parameters of the NN could also be safety relevant and the function performed by AI must be validated in the overall system. A safety qualification period will usually also be necessary within this context. The requirements of the SIL with regard to software should be easy to fulfil, since in principle only a relatively easy formula has to be implemented.
Secondly, the AI itself is not deterministic, but it makes decisions with certain probabilities. Therefore, it makes sense to assign part of the TFFR to the AI, i.e. the technical system's dangerous failure rate, see [3] and [2].
The first part of the safety case is relatively easy to perform, while the second becomes more difficult.

5.1 The analytical approach

There are interesting issues involved in the validation of AI: for example, if an error occurs in the validation of the AI, then the same error will occur again during its use, if the same data set occurs, or the error will occur with high probability, if a "similar" data set occurs. This could be interpreted as a systematic error in the AI function. Even if the fraction of errors is very small, is it permissible to commission a safety-critical system that is known to react incorrectly to certain data sets? Or must these data sets be exempted along with their "environment"? Or is it even necessary to require that the validation is error-free? However, this could cause problems if errors occur during safety qualification testing or operations. This means that one possibly has to tolerate errors and thus a certain hazard rate that arises due to erroneous decisions made by the AI function.
When verifying the TFFR, it is necessary to take into account the fact that the requirements are usually so strict that they can only be verified for low safety levels, e.g. SIL1, using practical verification, e.g. tested or proven in use arguments. For higher SIL, analytical proof will usually be necessary. If one looks at how such verifications are carried out in other areas, e.g. for computer architectures or data transmission, then one realises that such verifications are only successful if a sensible model is found. This insight is not new, as even the Turing laureate Pearl [15] has already concluded that it is necessary to "*formulate a model of the process that generates the data, or at least some aspects of that process*". It is necessary to understand that that is, at least in part, what the structure or properties of the data are. Only then can one expect to be able to derive properties from the (partial) model by means of analytical results that meet the requirements for a safety case for higher SILs. In practical terms, this means that such a proof will be easier for data that has a certain structure, e.g. due to the physical properties of the data sources, as may be the case for sensor data. On the other hand, from the perspective of the methods, one can expect that a safety case will be easier the more one understands the structure of AI methods. NN are a difficult example here, as they sometimes produce amazing results, but at the same time are difficult to explain. It could be that other AI methods, such as Support Vector Machines (SVM) [1], have advantages in establishing a safety case since they have a more explicable structure and provable properties. In any case, a "model-oriented" approach to the safety case would have the advantage that it could be conducted on the basis of today's established safety standards such as [14] or [16], and would not require any additional lengthy standardisation process as started with [6] or [17].

Sensordaten der Fall sein kann. Auf der anderen Seite, der Methoden-Seite, kann man erwarten, dass ein Sicherheitsnachweis einfacher sein wird, je mehr man die Struktur der KI-Methoden versteht. Hier sind NN ein schwieriges Beispiel, da sie zwar teilweise verblüffende Ergebnisse liefern, aber gleichzeitig schwer erklärbar sind. Es könnte sein, dass andere KI-Methoden, wie z. B. Support Vector Machines (SVM) [1], Vorteile bei der Sicherheitsnachweisführung haben, da sie eine besser erklärbare Struktur und nachweisbare Eigenschaften besitzen.

In jedem Fall besäße ein „modellorientierter" Ansatz für den Sicherheitsnachweis den Vorteil, dass er auf Grundlage der heute etablierten Sicherheitsstandards wie [14] oder [16] führbar wäre und es keiner langwierigen zusätzlichen Standardisierung bedarf, wie sie mit [6] oder [17] begonnen wurde.

5.2 Statistischer Ansatz

Alternativ zum analytischen Ansatz ist es auch möglich, einen statistischen Ansatz zu wählen. Man würde die KI dann als Black Box ansehen. Diese würde in einem ersten Schritt trainiert und danach würde der Zustand des KI-Systems eingefroren.

Nun beginnt die Phase der Verifizierung und des statistischen Testens. Die KI würde nun mit einer sehr großen Stichprobe von Daten konfrontiert, wobei neben den Daten selbst auch bekannt ist, welche Entscheidung von der KI jeweils zu erwarten ist. Man kann hierzu Daten simulieren oder auch die KI einfach mitlaufen lassen, wobei sie in dieser Testphase noch keine Entscheidungen trifft.

5.2 The statistical approach

It is also possible to take a statistical approach as an alternative to the analytical approach. The AI would then be regarded as a black box. The AI would be trained in the first step and then the state of the AI system would be frozen.

Now the verification and statistical testing phase begins. The AI would now be confronted with a very large data sample, whereby it is known which decision should be expected from the AI in each case in addition to the data itself. One can simulate data for this purpose or simply let the AI run along in an idle mode, where it does not yet make any decisions in this test phase.

If the AI has been statistically and representatively tested for a long enough time, it is possible to use an approach equivalent to one proven in use; for details see [18]. The mathematics of this are quite simple. If we are talking about a system that is supposed to make, say, classification-like binary decisions, then we can calculate from a test sample of size n, where the system has always made correct decisions, that the probability of a wrong decision is 3/n with 95 % statistical confidence.

The problems, however, lie elsewhere.

1. The test sample size
   On the one hand, a sufficiently large sample size must be generated. This can be difficult, if the events to be observed do not occur very frequently. It is also possible to use bootstrap procedures, i.e. to reproduce further random samples based on the already drawn sample; see [19]. If only the original data is used,

Wenn die KI lange genug statistisch und repräsentativ getestet wurde, kann man einen Ansatz verwenden, der dem der Betriebsbewährung äquivalent ist, für Details hierzu siehe [18]. Die Mathematik hierzu ist recht einfach. Wenn es um ein System geht, das z. B. klassifikationsähnliche binäre Entscheidungen treffen soll, dann kann man aus einer Teststichprobe vom Umfang n, bei der das System stets richtige Entscheidungen getroffen hat, bei einer statistischen Sicherheit von 95 % berechnen, dass die Wahrscheinlichkeit einer Fehlentscheidung 3/n ist.

Die Probleme liegen jedoch an anderer Stelle.

1. Stichprobenumfang für das Testen
   Einerseits ist ein hinreichend großer Stichprobenumfang zu generieren. Das kann schwierig sein, wenn die zu beobachtenden Ereignisse nicht sehr häufig eintreten. Man kann sich auch mit Bootstrap-Verfahren behelfen, d. h. man reproduziert aus der gezogenen Stichprobe wiederum zufällige weitere Stichproben, siehe z. B. [19]. Wenn man dabei einfach die Urdaten verwendet, reproduziert man genau genommen nur die ursprüngliche Stichprobe und hat wenig neuen Erkenntnisgewinn. Es gibt auch die Möglichkeit, an die beobachteten Daten, zumindest teilweise, ein Modell anzupassen und von diesem wiederum zu simulieren [20]. Die Schwäche dieses Vorgehens ist jedoch, dass das Modell geeignet sein muss. Ein weiterer Weg ist, die Komplexität des Problems zu nutzen. Die Daten sind in vielen Fällen pro Element der Stichprobe sehr umfangreich, d. h. man kann die Daten zerlegen und neu kombinieren. Wenn die Daten bespielweise aus Sequenzen von Sensordaten bestehen, die ein bestimmtes Objekt beschreiben, dann ist es möglich jeweils Teile dieser Daten neu zu kombinieren, und man erhielte damit die Beschreibung eines neuen Objektes. Zusammenfassend kann man sagen, dass eine Simulation durchaus geeignet ist, man muss nur einen guten Ansatz haben und dafür sorgen, dass man nicht nur die ursprüngliche Stichprobe einfach aufbläht und nur nominell größer macht, ohne neue Fälle hineinzubringen.
2. Repräsentativität der Stichprobe für das Testen
   Statistisch gesehen muss die Stichprobe repräsentativ sein. Das bedeutet, dass zunächst die KI auf alle möglichen zukünftig vorkommenden Daten angelernt werden muss und danach auch mit derartigen Daten getestet werden muss. Die Herausforderung hierbei ist, dass die jeweiligen Stichproben der Realität entsprechen und alle relevanten Fälle abdecken. Man kann sich vorstellen, dass gerade bei z. B. Messwerten hier sehr viel Arbeit dahinter steckt. In bestimmten Fällen muss man dazu die Daten sehr ausführlich analysieren, sodass man ggf. eher ein gutes statistisches Model erzeugt hat, als eine KI angelernt und nachweislich getestet hat.

Trotzdem halten die Autoren diese Herangehensweise für machbar, insbesondere, wenn man die Daten relativ einfach automatisch erheben kann – dann kommt man leicht zu einem großen Stichprobenumfang, und man kann sich auch anhand einiger relevanter Parameter überlegen, ob die Stichprobe repräsentativ ist [21].

6 Diskussion und Zusammenfassung

In diesem Beitrag wurde versucht, an einem gut bekannten Analogiebeispiel möglichst allgemeinverständlich herauszuarbeiten, welche Herausforderungen bewältigt werden müssen, bevor ein KI-basiertes System in der Eisenbahntechnik autonom sicherheitsrelevante Entscheidungen treffen dürfte. Aufgrund des beispielhaften Charakters der Ableitung kann kein Anspruch auf Vollständigkeit erhoben werden, allerdings sind diese allgemeinen Anforderungen mindestens zu erfüllen:

it is possible to reproduce the original sample and not gain much new knowledge. There is also the possibility of fitting a model to the observed data, at least in part, and simulating from that in turn [20]. The weakness of this approach, however, lies in the fact that the model must be suited. Another way involves taking advantage of the complexity of the problem. In many cases, the data is very large for each element of the sample, i.e. one can decompose and recombine the data. For example, if the data consists of sensor data sequences describing a certain object, it is possible to recombine parts of this data and thus obtain the description of a new object. In summary, it can be said that a simulation is quite suitable, but it is necessary to have a good approach and to make sure that one has not merely inflated the original sample and made it only nominally larger without bringing in any new test cases.

2. The representativeness of the test sample
   The sample must be statistically representative. This means that the AI must first be trained for all possible future data and then also be tested with such data. The challenge here is to ensure that the respective samples correspond to reality and cover all the relevant cases. One can imagine that this involves a lot of work especially with the measured values, for example. In certain cases, the data must be analysed in great detail, so it might be easier to derive a good statistical model rather than to arrive at AI that has been trained and demonstrably tested.

Nevertheless, the authors consider this approach to be feasible, especially if it is relatively easy to collect the data automatically - then it is possible to easily arrive at a large sample size and to also consider whether the sample is representative based on some relevant parameters [21].

6 Discussion and summary

This paper has attempted to work out, using a well-known analogy example and in the most general possible manner, which challenges have to be overcome before an AI-based system in railway engineering is allowed to make autonomous safety-relevant decisions. Due to the paradigmatic nature of the reflections, no claim can be made for completeness. However, the following general requirements must at least be met:

- the data used to train and validate the AI must be representative for the operating environment.
- there must be clear rules on how to deal with anomalies or "outliers" in the data
- the operating environment must not change.
- the AI approach must be flexible enough to approximate the (unknown) "true" decision function well enough, but also to avoid overfitting.

In addition, the following minimum requirements also arise when using the CENELEC standards:

- the AI function must have been validated with the system, where it will be used.
- there must have been a safety qualification testing period.
- the validation data must not have been used to train the AI function.
- the SIL requirements must have been proven for the implementation of the AI function, i.e. for the used software and, if applicable, tools.
- a tolerable functional failure rate (TFFR) must have been demonstrated, either by means of a test or operational experience (for a low SIL) or using an analytical model (for a higher SIL).

- Die Daten, anhand derer die KI trainiert und validiert wird, müssen repräsentativ für die Einsatzumgebung sein.
- Es muss klare Regeln geben, wie mit Anomalien oder „Ausreißern" in den Daten verfahren wird.
- Die Einsatzumgebung darf sich nicht verändern.
- Der KI-Ansatz muss flexibel genug sein, um die (unbekannte) „wahre" Entscheidungsfunktion gut genug anzunähern, aber auch Overfitting vermeiden.

Zusätzlich ergeben sich bei Anwendung der CENELEC-Normen mindestens diese Anforderungen:

- Die KI-Funktion muss mit dem System, in dem sie eingesetzt wird, validiert werden.
- Es muss eine Sicherheitserprobung stattfinden.
- Die Validierungs-Daten dürfen nicht zum Training der KI-Funktion verwendet werden.
- Für die Implementierung der KI-Funktion müssen die SIL-Anforderungen z. B. für die Software und ggf. die eingesetzten Tools nachgewiesen werden.
- Es muss eine Gefährdungsrate (TFFR) nachgewiesen werden, entweder durch Test- oder Betriebserfahrung (für niedrige SIL) oder durch ein analytisches Modell (für höhere SIL).

Es hat sich gezeigt, dass es hilfreich ist, erst einmal anscheinend einfache Probleme zu lösen, bevor man sich komplizierteren Herausforderungen stellt. Derartige Probleme wurden im Rahmen der „AI Dependability Assessment Challenge" von Studenten und Doktoranden bearbeitet [23]. Genauso könnte es schneller zum Ziel führen, wenn man versucht, die Probleme auf Grundlage der heute gültigen Standards zu lösen, statt neue Normen zu fordern. Dann kann man auch die üblichen Verfahren der Sicherheitsbegutachtung [22] anwenden. ■

It has been shown that it can be helpful to first solve the apparently simple problems before taking on any more complicated challenges. Such problems have been dealt with in the "AI Dependability Assessment Challenge" by students and PhD students [23]. In the same way, it might be quicker to reach the goal, if one attempts to solve the problems based on the standards that are currently in force, rather than demanding new standards. Then the usual safety assessment procedures [22] can also be applied. ■

AUTOREN | AUTHORS

Dr. Hendrik Schäbe
Principal Assessor RAMS
TÜV Rheinland Intertraffic GmbH Köln
Anschrift / *Address*: Am Grauen Stein, D-51105 Köln
E-Mail: schaebe@de.tuv.com

Prof. Jens Braband
Principal Key Expert
Siemens Mobility GmbH
Anschrift / *Address*: Ackerstraße 22, D-38126 Braunschweig
E-Mail: jens.braband@siemens.com

LITERATUR | LITERATURE

[1] Russell, S.; Norvig, P.: Künstliche Intelligenz, Pearson, 2012
[2] Hemzal. G.; Strobel, T.; Großmann, J.; Schlingloff, B.-H.; Leuschel, M.; Sadeghipour, S.; Firnkörn, J.: KI-LOK – A joint test procedure project for AI-based components used in railway operations, SIGNAL+DRAHT (113) 10/2021, S. 6 – 15
[3] Braband, J.; Schäbe, H.: On Safety Assessment of Artificial Intelligence, Dependability, vol. 20, No. 4, S. 25-35, 2020
[4] Brunette, E.S.; Flemmer, R.C.; Flemmer, C.L. (2009): A Review of artificial Intelligence, Proc. 4th International Conference on Autonomous Robots and Agents, Feb. 10-12, 2009, Wellington, p. 385-392
[5] DIN/DKE: Normungsroadmap Künstliche Intelligenz, 2020
[6] Underwriter Laboratories: Standard or the Evaluation of Autonomous Products, UL 4600, 2020
[7] Putzer, H. (2019): Ein strukturierter Ansatz für funktional sichere KI, Presentation at DKE Funktionale Sicherheit, Erfurt
[8] Berghof, C.; Biggio, B.; Brummel, E.; Danos, V.; Foms, T.; Ehrich, H.; Gantevoort, T.; Hammer, B.; Iden, J.; Jacob, S.; Khlaaf, H.; Komrowski, L.; Kröwing, R.; Metzen, J.H.; Neu, M.; Petsch, F.; Poretschkin, M.; Samek, W.; Schäbe, H.; von Twickel, A.; Vehev, M.; Wiegand, T.: Towards Auditable AI systems, White Paper TÜV Verband, BSI, Fraunhofer, Mai 2021
[9] Vapnik, V.N.: The Nature of Statistical Learning Theory, Springer, 2010
[10] Vapnik, V.N.; Chervonenkis, A. Ja.: Theorie der Zeichenerkennung, Nauka, Moskau 1974 (in Russisch)
[11] Anscombe, F. J. (1973): Graphs in Statistical Analysis, American Statistician, 27 (1): 17–21
[12] Cybenko, G. (1989): Approximations by superpositions of sigmoidal functions: Mathematics of Control, Signals and Systems, 2(4), 303–314
[13] Schäbe, H. (2018): SIL Apportionment and SIL Allocation in: Handbook of RAMS in Railway systems – Theory and Practice, Mahboob, Q.; Zio, E. (Eds.), Boca Raton, Taylor and Francis, chapter 5, p. 69-78
[14] DIN EN 50129: Sicherheitsbezogene elektronische Systeme für Signaltechnik, 2019
[15] Pearl, J.; Mackenzie, D. (2018): The Book of Why, Penguin Science
[16] IEC 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems, 2010
[17] VDE: Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen, VDE-AR-E 2842-61-6, Anwendungsregel, 2020
[18] Braband, J.; Gall, H.; Schäbe, H. (2018): Proven in Use for Software: Assigning an SIL Based on Statistics in: Handbook of RAMS in Railway systems – Theory and Practice, Mahboob, Q.; Zio, E. (Eds.), 2018, Boca Raton, Taylor and Francis, Chapter 19, p. 337-350
[19] Davison, A.C.; Hinkley, D.V.: Bootstrap methods and their application, Cambridge University Press 1997
[20] Didona D.; Romano, R.: Using Analytical Models to Bootstrap Machine Learning Performance Predictors, 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), 10 p.
[21] Strom, R.: Wahrscheinlichkeitsrechnung, Mathematische Statistik, Qualitätskontrolle, VEB Fachbuchverlag Leipzig, 1988
[22] Wigger, P. (2018): Independent Safety Assessment – Process and Methodology in: Handbook of RAMS in Railway systems – Theory and Practice, Mahboob, Q.; Zio, E. (Eds.), Boca Raton, Taylor and Francis, chapter 5, p. 475-485
[23] Siemens Mobility GmbH: AI-DA Challenge, 2021, https://ecosystem.siemens.com/universityrelations/ai-da-challenge-ai-dependability-assessment/overview
[24] Braband, J.: Künstliche Intelligenz – Mit Sicherheit?, Deine Bahn, Nr. 4/2021, S. 30-35