Sachbericht zum Verwendungsnachweis des WIR!-Projekts

KI-bezogene Test- und Zulassungsmethoden (SRCC-KI)

(FKZ 03WIR1205)

Zuwendungsempfänger: **Technische Universität Chemnitz**

Laufzeit: **08/2020 – 01/2022**

Projektleiterin: Prof. Dr. Dagmar Gesmann-Nuissl

Teil I – Kurzbericht

1 Ursprüngliche Aufgabenstellung

Zukünftig werden viele Innovationen im Feld des intelligenten Schienenverkehrs Methoden der künstlichen Intelligenz (KI) nutzen. Allerdings sind die üblichen Verfahren für Sicherheitsnachweise für Systeme auf KI-Basis aktuell langwierig bzw. ohne Betrieb überhaupt nicht oder nur in sehr eingeschränkten Grenzen anwendbar. Ziel des Vorhabens war es daher zum einen, einen Anwenderkompass für den Erprobungs- und Zulassungsprozess im Schienenverkehr zu erarbeiten, um Forscher, Entwickler, Hersteller und Inverkehrbringer bahntechnischer Innovationen zu unterstützen, die sich erstmals oder selten mit bahntechnischen Innovationen beschäftigen. Zum anderen diente das Vorhaben dazu, mögliche Wege zur Sicherheitsnachweisführung für KI-Verfahren zu identifizieren und Bedarfe zur Anpassung bzw. Überarbeitung der geltenden Zulassungsvorschriften abzuleiten. Der Fokus lag hier auf der Zulassung von KI für sicherheitsrelevante Anwendungen im Schienenverkehr.

Um die vorgenannten Projektziele greifbarer zu machen, sollte sich das Forschungsvorhaben auf zwei konkrete bahnspezifische Beispiele beziehen, denen eine besondere praktische Relevanz in Verbindung mit dem Einsatz Künstlicher Intelligenz zukommt. Die Projektpartner verständigten sich auf die automatisierte Fahrtwegerkennung (Hinderniserkennung) sowie das Fiber-Optic-Sensing zur genauen Positionsbestimmung eines Zuges.

Das Forschungsvorhaben wurde an der TU Chemnitz durch die Professuren Medieninformatik (Prof. Dr. Maximilian Eibl), Betriebssysteme (Prof. Dr. Matthias Werner) sowie Privatrecht und Recht des geistigen Eigentums (Prof. Dr. Dagmar Gesmann-Nuissl) und dem Zentrum für Wissens- und Technologietransfer der TU Chemnitz (ZWT) durchgeführt. Über eine Auftragsvergabe waren zudem das Institut für Verkehrsforschung des Deutschen Zentrums für Luft- und Raumfahrt e. V. (DLR), die Ingenieurgesellschaft für Verkehrssicherungstechnik GmbH (IVS), das Fraunhofer-Institut für Werkzeugmaschinen und Umformtechnik (IWU), das Fraunhofer-Institut für Elektronische Nanosysteme (ENAS), die Siemens Mobility GmbH, die IFB Institut für Bahntechnik GmbH, die IAV GmbH Ingenieurgesellschaft Auto und Verkehr sowie der Smart Rail Connectivity Campus e. V. eingebunden. Die Projektleitung nahm der Lehrstuhl Privatrecht und Recht des geistigen Eigentums (Prof. Dr. Dagmar Gesmann-Nuissl) wahr.

2 Ablauf des Vorhabens

Das Projektvorhaben gliederte sich in sieben ineinander greifende Arbeitspakete – der strukturelle und inhaltliche Zusammenhang ist in Abbildung 1 dargestellt.

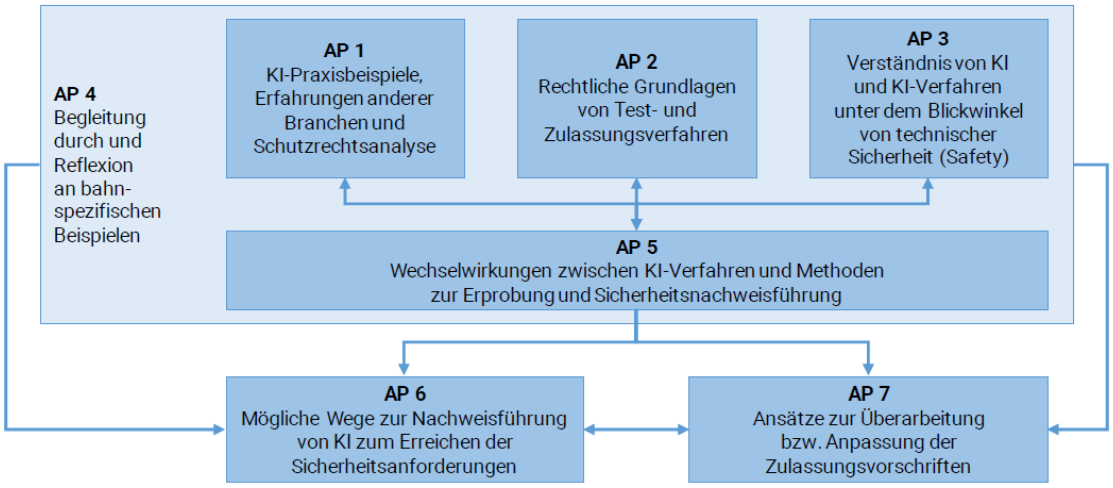


Abbildung 1: Arbeitspakete des Vorhabens

Für die Bearbeitung dieser Arbeitspakete waren folgende Projektbeteiligte verantwortlich:

AP 1	**TU Chemnitz (Professur Medieninformatik**/Professur Betriebssysteme)
AP 2	**TU Chemnitz (Professur Privatrecht und Recht des geistigen Eigentums)**
AP 3	**TU Chemnitz (Professur Betriebssysteme**/Professur Medieninformatik)
AP 4	**DLR**
AP 5	**TU Chemnitz (ZWT)**
AP 6	**IVS**
AP 7	**TU Chemnitz (Professur Privatrecht und Recht des geistigen Eigentums)**

Die gewonnenen Erkenntnisse der einzelnen Arbeitspakete wurden untereinander ausgetauscht und mit den weiteren Projektbeteiligten in regelmäßigen Treffen evaluiert.

Hinsichtlich des Standes von Wissenschaft und Technik bezog sich das Vorhaben auf die vorgefundene Rechts- bzw. Normenlage. Die KI-Methoden und -Verfahren setzten auf den gesicherten und eingeführten Stand von Wissenschaft und Technik auf. Für die Darstellung des existierenden Zulassungsverfahren im Bahnsektor – national und europäisch – konnte auf diverse Veröffentlichungen[1] zurückgegriffen werden, die bedarfsgerecht und projektbezogen aufbereitet wurden. Konkrete Vorarbeiten zur Neujustierung eines KI-bezogenen Test- und Zulassungsverfahrens im Bahnsektor waren bis Projektende noch nicht vorhanden; Projektvorhaben mit ähnlichen Schwerpunkten (z. B. das Forschungsprojekt ATO-Risk) werden erst Ende 2023 abgeschlossen.

3 Zusammenfassung der wesentlichen Ergebnisse

Das System der Eisenbahnzulassung ist durchaus funktional und in seiner Grundsystematik den technischen Herausforderungen der Zukunft gewachsen. Für die Entwicklung einer sicherheitsrelevanten KI-Anwendung gilt der Systementwicklungsprozess nach EN 50126 und EN 50129. Für die Softwareentwicklung gelten zusätzlich die Anforderungen an den Entwicklungsprozess gemäß EN 50128. Keine dieser Normen berücksichtigt derzeit allerdings den Einsatz von KI.

Aus diesem Grund wurde überprüft, inwieweit die Entwicklung einer sicherheitsrelevanten KI-Anwendung auf den normativen Entwicklungsprozess abgebildet werden kann. Es wurde dargestellt, wie die wesentlichen normativ geforderten Inhalte der einzelnen Phasen auf eine KI-Anwendung übertragen werden können und welche Fragestellungen bzw. offenen Punkte gegenwärtig noch nicht hinreichend beantwortet sind.

Diesbezüglich konnten verschiedene Defizite aufgedeckt werden: In formaler Hinsicht ist zunächst darauf hinzuweisen, dass bei der Formulierung und Übersetzung der einschlägigen europäischen Normen erheblicher Verbesserungsbedarf besteht. Sie sind teilweise inkohärent gefasst und uneinheitlich übersetzt, was insbesondere die Anwender irritieren und Fehlinterpretationen hervorrufen kann. Bei Einordnung der sicherheitsrelevanten KI-Anwendung in den normativen Rahmen konnte aufgezeigt werden, dass der Nachweis der funktionalen Sicherheit maßgeblich mit der Erklärbarkeit von KI korreliert. Hier wurde nach Konzepten gesucht, die das Manko der Erklärbarkeit überwinden helfen (u. a. weiter aufbauend auf bereits bestehende Konzepte wie bspw. LIME oder Tree Regularization), sowie der Frage nachgegangen, inwieweit man die Erklärbarkeit zu Teilen verzichtbar stellen könnte (unter einer Akzeptanz von Restrisiken).

[1] Hierzu gehörten u. a. Fendrich, L.; Fengler, W. (Hrsg.): Handbuch Eisenbahninfrastruktur, 3. Auflage, Berlin 2019, S. 981 ff.; Salander, C.: Das Europäische Bahnsystem: Akteure – Prozesse – Regelwerke, Berlin 2019, S. 85 ff.

Teil II – Ausführlicher Bericht

1 Durchgeführte Arbeiten

Das Projektvorhaben gliederte sich in sieben ineinander greifende Arbeitspakete – der strukturelle und inhaltliche Zusammenhang ist in Abbildung 1 dargestellt.

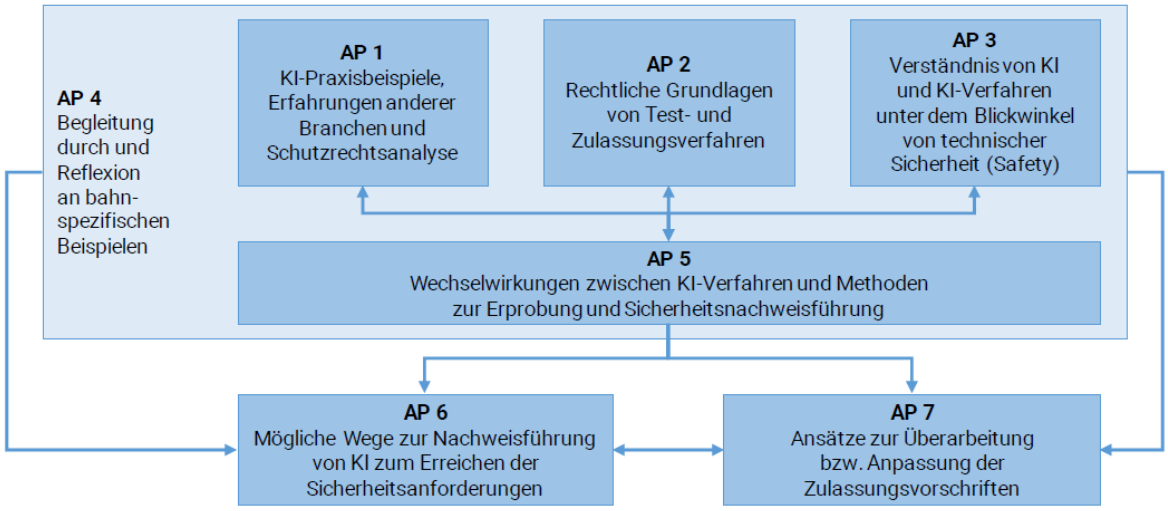


Abbildung 2: Arbeitspakete des Vorhabens

Die Arbeitspakete 1 bis 3 sind dabei als projektspezifische Grundlagenarbeiten zu verstehen, auf denen die weiteren Arbeitspakete aufbauen. Das Arbeitspaket 4 begleitet diese Arbeiten, indem zur Begrenzung der Komplexität ausgewählte bahnspezifische Beispiele, die einen möglichst großen Lösungsraum für potenzielle KI-Anwendungen aufspannen, beschrieben und die Arbeiten daran ausgerichtet werden. Den Kern des Vorhabens bilden die Arbeitspakete 5 bis 7, in denen die Auswirkungen von KI-basierten Innovationen auf Test- und Zulassungsmethoden untersucht und Rückschlüsse auf Wege zur Nachweisführung sowie auf Anpassungsbedarfe der aktuellen Zulassungsvorschriften gezogen werden. Dabei erstreckte sich das Projekt über eine Laufzeit von 18 Monaten im Zeitraum zwischen August 2020 und Januar 2022.

AP 1 KI-Praxisbeispiele, Erfahrungen anderer Branchen und Schutzrechtsanalyse

In AP 1 sollten mittels Literatur-/Patentrecherche und technischer Evaluation ein umfassender Überblick über den aktuellen Stand der Technik im Bereich des autonomen Fahrens mit KI-gestützten Systemen gegeben, ein Testsystem zur Untersuchung von Angriffsvektoren und nicht-deterministischem Verhalten auf Basis veröffentlichter Daten- und Code-Repositories erstellt sowie ein systematischer Überblick über Angriffsvektoren und deren Auswirkungen auf KI-gestützte Sicherheitssysteme sowie deren potentielle Bedeutung im Bahnverkehr erarbeitet werden.

Im Rahmen der Bearbeitung von AP 1 wurde eine Recherche zu möglichen unerwünschten Verhaltensweisen, mit dem Schwerpunkt auf sog. Adversarial Attacks, von Neuronalen Netzen durchgeführt. Im Ergebnis entstand ein Überblick über die Gefahren und Angriffsvektoren, die durch diese neuen Techniken entstehen. So können Systeme angegriffen werden, ohne dass der Angreifer die inneren Strukturen des Zielsystems kennt (Blackbox), da Adversarial Attacks bei einer Reihe von Neuronalen Netzen ähnliche Auswirkungen aufweisen. Es wurden Counter-Maßnahmen recherchiert, die versuchen, dieses Problem zu adressieren, indem der Trainingsvorgang der Neuronalen Netze angepasst wird. Während der Recherche wurde ein Testsystem zur Untersuchung von Adversarial Attacks implementiert.

Um neuronale Netze in irgendeiner Form nachweisen zu können und dadurch den Vorgang der Verifikation zu vereinfachen, wurde darüber hinaus eine Literaturrecherche zum Thema Nachweisbarkeit von tiefen Neuronalen Netzen vorgenommen. Diese zeigte, dass bereits Forschungsgebiete mit dem Ziel existieren, KI für menschliches Verstehen aufzubereiten. Der Bereich der „Explainable AI“ versucht, die Ausgaben von KI zu interpretieren bzw. zu zeigen, wieso das Neuronale Netz zu diesem Ergebnis gekommen ist. Im Vorhaben wurde ein Überblick der in der Literatur gefundenen Möglichkeiten erstellt, u. a. Techniken zum Aufzeigen der für die KI zur Entscheidung beitragenden Bereiche innerhalb der Eingabe. Eine Methodik zur Nachweisbarkeit, die sog. Tree-Regularization, wurde dahingehend genauer untersucht, ob eine Beweislegung über die angelernten „Features“ des Neuronalen Netzes möglich ist. Nach einer umfassenden Recherche wurde diese Möglichkeit aufgrund nicht vorhandener Publikationen verworfen. Die Literaturrecherche zur „Tree-Regularization“ ergab vor allem zwei Veröffentlichungen, welche zwar erfolgreich über ein Neuronales Netz einen beweisbaren Entscheidungsbaum trainieren, jedoch v. a. im medizinischen Sektor angewandt werden. Es wurde daher angekommen, dass Entscheidungsbäume nicht ausreichend komplex sein können, um die ausgewählten bahnspezifischen Beispiele abbilden zu können.

Zudem wurde ein Überblick zu Techniken der Verifikation/Validierung von KIs angefertigt, der allerdings nicht speziell für autonomes Fahren, sondern allgemein für KI-Systeme gilt.

AP 2 Rechtliche Grundlagen von Test- und Zulassungsverfahren

Im AP 2 sollten mit den Methoden der Literatur- und Normenrecherche sowie durch rechtsvergleichende Analysen ein systematischer Überblick über Test- und Zulassungsverfahren im Bahnsektor inkl. einer abstrakten Beschreibung der Test- und Zulassungsprozesse sowie ein systematischer Überblick über Test- und Zulassungsverfahren für sicherheitskritische Anwendungen und für produktbezogene KI-Anwendungen außerhalb des Bahnsektors erstellt werden. Zudem sollten die Lücken/Defizite der aktuellen Test- und Zulassungsprozesse im Hinblick auf die möglichen KI-Anwendungsbeispiele aufgezeigt werden.

Im *AP 2.1* wurden die Zulassungssysteme im Eisenbahnbereich untersucht, indem die Zulassungsregelungen für mobile Teilsysteme sowie diejenigen für streckenseitige Teilsysteme betrachtet wurden. Zudem wurde die generelle Funktionsweise von Systemen künstlicher Intelligenz in ihrem rechtlichen Rahmen untersucht.

Die Zulassungsregelungen sowohl für mobile als auch für streckenseitige Systeme sind dabei stark fragmentiert. Während die Zulassung der mobilen Systeme weitgehend unionsrechtlich organisiert und diesbezüglich vereinheitlicht ist, ist die Zulassung der streckenseitigen Teilsysteme den Mitgliedsstaaten vorbehalten. Dabei haben diese die Kompatibilität zu den mobilen Teilsystemen zu beachten. Beide haben jedoch die generelle Genehmigungspflicht für die Inbetriebnahme von Teilesystemen der jeweiligen Systeme gemeinsam.

Für die Zulassung der mobilen Teilsysteme ist in ihrer konkreten Ausgestaltung vor allem die Durchführungsverordnung (EU) 2018/545 maßgeblich, die die Durchführung der Richtlinie (EU) 2016/797 konkretisiert und in das nationale Zulassungsregime über das Allgemeine Eisenbahngesetz und die Eisenbahninbetriebnahmegenehmigungsverordnung integriert ist. Sie ist dergestalt aufgebaut, dass lediglich eine Plausibilitätsprüfung stattfindet. Dafür sind die Prüfberichte der Komponenten der einzelnen Bauteile/Systeme des mobilen Systems bei der jeweiligen Prüfstelle einzureichen. Es wird geprüft, ob diese vollständig und plausibel sind. Prüfstelle kann nach Verwendungsgebiet die ERA oder eine nationale Sicherheitsbehörde (NSB) sein. Soweit das mobile System im Netz mehr als eines Mitgliedsstaates genutzt werden soll, muss die Prüfung

durch die ERA erfolgen, bei Beschränkung der Nutzung auf das Netz eben eines Mitgliedsstaates kann nach Wahl des Antragstellers auch die NSB als Genehmigungsbehörde fungieren.

Im Hinblick auf die Zulassung von KI-Anwendungen ergibt sich das Problem, dass die Anwendungen im derzeitigen Zulassungsregime durch Zertifizierungsstellen überprüft werden müssen, für eine derartige Überprüfung jedoch noch keine Normen bestehen. Auch herrscht bisher Unklarheit darüber, wie ein entsprechender Sicherheitsnachweis zu führen sein könnte, da nach Stand der Forschung noch kein Nachweis darüber geführt werden kann, wie eine einzusetzende KI (in den Anwendungsfällen wird von Deep-Neural-Networks ausgegangen) ihre Entscheidungen trifft.

Im *AP 2.2* wurden im Rahmen der Vorhabenbeschreibung Test- und Zulassungsverfahren außerhalb des Bahnsektors erforscht. Das Augenmerk lag auf dem Automotive und Aviation Sektor.

§ 1a Abs. 3 StVG regelt in Verbindung mit internationalen Vorschriften die Zulassung von Fahrzeugen mit hoch- und vollautomatisierten Fahrfunktionen. Unter diesen internationalen Vorschriften sind u. a. die UNECE-Regelungen zu verstehen. Grundsätzlich ist in den einschlägigen UNECE-Regelungen auch die Verwendung von KI-Systemen impliziert (vgl. etwa 3.2. Annex 4 UNECE Nr. 157 für das Spurhalteassistenzsystem). Selbstverständlich müssen die verwendeten Systeme den Anforderungen an die Sicherheit entsprechen. Maßstab für Sicherheit im Automotive-Bereich ist grundsätzlich der Vergleich mit einem manuell gesteuerten Fahrzeug (vgl. etwa 3.4.4. UNECE Nr. 157). Für den Sicherheitsnachweis spielt die Dokumentation durch den Hersteller eine zentrale Rolle, die durch die Genehmigungsbehörde geprüft wird. Maßstab für die Überprüfung sind die technischen Standards ISO 26262 und SOTIF ISO/PAS 21448 (7. Anhang 4 UNECE Nr. 157).[2] Im Jahr 2021 wurde darüber hinaus die ISO/TR 4804 veröffentlicht, die auf die Sicherheit innovativer Technologien Bezug nimmt und in Anhang B ausdrücklich Deep Learning behandelt. In den genannten Regelungen wird auf diesen Standard aber bisher noch nicht explizit Bezug genommen.

Bei den Regelungen der oben genannten Fahrerassistenzsysteme wird jedoch immer von einem im Fahrzeug anwesenden menschlichen Fahrzeugführer ausgegangen. Soll für den Rail-Bereich der Triebfahrzeugführer vollständig durch ein KI-System ersetzt werden, bietet sich ein Vergleich mit den neuen nationalen Regelungen zur Zulassung von „Kraftfahrzeugen mit autonomer Fahrfunktion in festgelegten Betriebsbereichen“ an (Zulassung § 1e Abs. 1 StVG). Ein Paradigmenwechsel hinsichtlich der hohen Bedeutung des mittels industrieseitigen Standards ausgearbeiteten Sicherheitskonzeptes wird nach aktuellen Erkenntnissen auch mit der Durchführungsverordnung zum autonomen Fahren (AFGBV – Autonome Fahrzeug-Genehmigungs- und Betriebsverordnung) nicht eintreten (Anlage 1 Teil 1 Nr. 7 bzw. Anlage 4 AFGBV). Für technische Einzelheiten insbesondere auch hinsichtlich der Sicherheit wird in der AFGBV u. a. auch auf die existierenden UNECE-Regelungen verwiesen. Auch soll das Kraftfahrt-Bundesamt auf bewährte Weise das Sicherheitskonzept anhand der vom Fahrzeughersteller durchgeführten Dokumentation prüfen. Die Sicherheit soll weiterhin durch die „technische Aufsicht“ während der Fahrt in Echtzeit garantiert werden (§ 1d Abs. 3 StVG). Diese menschliche Aufsicht hat u. a. die Möglichkeit, die autonome Fahrfunktion zu deaktivieren, Fahrmanöver vorzuschlagen oder freizugeben (vgl. § 1e Abs. 2 StVG). Gänzlich ohne menschliche Einflussmöglichkeit wird das autonome Fahren auf deutschen Straßen im Regelverkehr damit auch nach den neuesten Regelungen noch nicht ermöglicht. Allerdings wurde für die Zukunft die Möglichkeit offen gelassen, dass die „technische Aufsicht“ mehr als ein Fahrzeug zugleich betreut.[3] Dies könnte evtl. auch für den Bahnverkehr einen gangbaren Zwischenschritt zum Sammeln von „Real-Word-Experience“ darstellen.

[2] Cyber-Security soll im Folgenden nicht näher betrachtet werden.
[3] BT-Drs. 19/27439, S. 29.

Für den Flugverkehr werden Safety-Aspekte unter dem Oberbegriff „Lufttüchtigkeit“ diskutiert. Neben Regelungen durch Verordnungen (Verordnung (EU) 2018/1139)[4] und Durchführungsverordnungen (VO (EU) Nr. 748/2012) werden technische Details sicherheitsrelevanter Flugzeugzertifizierung u. a. durch nicht unmittelbar[5] verbindliche technische Standards der EASA wie Zertifizierungspezifikationen (CS) und Accceptable Means of Comliance (AMC) kleinteilig geregelt. CS 25.1309 „Equipment, systems and installations“ definiert sicherheitstechnische Anforderungen an in Flugzeuge eingebaute Systeme. Das zugehörige Material zur Interpretationshilfe mit weiteren Richtlinien zur Risikoanalyse und -minimierung stellt AMC 25.1309 „System Design and Analyses“ dar. In Anhang II ist der Safety-Assessment Prozess überblicksartig dargestellt. Ergänzend wird auf weitere auch industrieseitige Standards AMC 20-115, SAE/ARP 4754A/EUROCAE ED-79A und SAE/ARP 4761 verwiesen. Als weitere wichtige Standards sind ARP4754A/ED-79A[6], EUROCAE ED-12C/RTCA DO-178C[7] sowie EUROCAE ED-76A/RTCA DO-200B[8] zu nennen. Alle diese Regelungen sind jedoch vor dem Hintergrund von konventionellen deterministischen Algorithmen entwickelt worden und weisen daher für KI-spezifische Risiken Regelungslücken auf.[9] Um diese Lücke zu schließen, haben die EASA und die Daedalean AG in einem gemeinsamen Projekt (Concepts of Design Assurance for Neural Networks – CoDANN)[10] die Zertifizierung von KI-getriebenen Flugsystemen vorbereitet. KI-Systeme sind für Zertifizierungszwecke in 3 Stufen eingeteilt. Aktuell ist ein erster Entwurf von Entwicklungsleitlinien für KI-Systeme der Stufe 1 veröffentlicht.[11] Künstliche Intelligenz auf Stufe 1 adressiert Systeme, die dem Menschen bei der Ausführung seiner Tätigkeiten assistieren. So erkennt KI in einem „runway detection system“ aus aufgenommenen Bildern das Vorhandensein und die Eckpunkte einer Landebahn und gibt diese Information an den Piloten weiter. Zertifizierungskonzepte für die weiteren KI-Stufen sind für die kommenden Jahre geplant. Daher kann diese sicherheitstechnische Zertifizierungshilfe nicht als Blaupause für die Ersetzung des Triebfahrzeugführers durch KI-Applikationen im Rail-Bereich herangezogen werden. Allerdings kann der gestaffelte Ansatz der Aviation-Branche eine Anregung zur schnelleren Zertifizierung von Anwendungen auch im Rail-Bereich sein. Ohne Zweifel betont er die hohen Anforderungen, welche an die Sicherheit von KI-Applikationen gestellt werden.

Insgesamt fällt auf, dass sowohl Automotive- als auch Aviation-Branche eine gestaffelte sicherheitstechnische Zertifizierung von KI anstreben. Für den Flugverkehr werden KI-Systeme je nach Leistung des Systems und Grad der Aufsicht durch den Menschen in verschiedene Stufen eingeteilt. Diese Stufen stellen steigende Anforderungen an den sicherheitstechnischen Zertifizierungsprozess. Für den Straßenverkehr wird die Verwendung von KI-Applikationen ohne direkte örtliche Anwesenheit eines menschlichen Fahrzeugführers zuerst in bestimmten Betriebsbereichen zugelassen. Weiterhin stellt die ‚technische Aufsicht‘ den reibungslosen Betrieb der autonomen Fahrzeuge sicher. Insgesamt werden hohe Anforderungen an die Sicherheit von KI-Applikationen gestellt. Mithilfe verschiedener Mittel soll die Sicherheit der Eingliederung von KI in den Realbetrieb garantiert werden. Diese generell absichernde Tendenz sollte auch für den Bahnverkehr richtungsweisend sein.

4 Insb. Anhang II ‚Grundlegende Anforderungen an die Lufttüchtigkeit`.
5 Sie zeichnen sich durch einen quasi-verbindlicher Charakter (Soft Law) aus.
6 The Guidelines for Development of Civil Aircraft and Systems.
7 The Software Considerations in Airborne Systems and Equipment Certification.
8 The Standards for Processing Aeronautical Data.
9 SAE Artificial Intelligence in Aeronautical Systems: Statement of Concerns AIR6988.
10 Zu dem Projekt sind neben dem ‚concept paper‘ der EASA auch die beiden technischen ‚reports‘ CoDANN I und CoDANN II in weiten Teilen veröffentlicht.
11 EASA, Concept Paper: First usable guidance for Level 1 machine learning applications – A deliverable of the EASA AI Roadmap.

AP 3 Verständnis von KI und KI-Verfahren unter dem Blickwinkel von technischer Sicherheit (Safety)

In AP 3 sollten mittels Literaturrecherche eine Liste der zu beachtenden Aspekte von KI-Ansätzen hinsichtlich technischer Sicherheit sowie eine Wissensmatrix von KI-Ansätzen bzgl. deren Aspekte hinsichtlich technischer Sicherheit erarbeitet werden. Da keine Klarheit über die Begrifflichkeiten auf dem Gebiet der künstlichen Intelligenz – insbesondere bzgl. des maschinellen Lernens und dessen Abgrenzung zu anderen KI-Ansätzen – und die Relevanz von algorithmischen Ansätzen (hier KI-Ansätzen) in Bezug auf die Systemverlässlichkeit und die funktionale Sicherheit gibt, war es die Aufgabe im AP 3, KI-Systeme aus dem Blickwinkel der Verlässlichkeit zu betrachten, wesentliche Begriffe zu definieren und dabei die kritischen Punkte herauszuarbeiten.

Ergebnis ist ein Report mit folgendem Inhalt: Der erste Teil des Reportes erörtert die Begriffe, Konzepte und ihre Beziehungen. Im Abschnitt 2 wird Verlässlichkeit aus Sicht der Forschung diskutiert. Abschnitt 3 diskutiert das Konzept der künstlichen Intelligenz, gibt eine grobe Taxonomie und erläutert grundsätzliche Ansätze für Lernverfahren in künstlichen neuronalen Netzen. Auf der Grundlage der Abschnitte 2 und 3 werden im Abschnitt 4 die Besonderheiten von Systemen der künstlichen Intelligenz in Bezug auf andere (sicherheitskritische) Systeme erarbeitet. Zudem wird auf die spezielle Rolle der Lernmusterdaten eingegangen. Dabei werden zwei Interpretationen für das Verständnis des Trainierens von Lernmusterdaten angeboten, die sich daraus ergebenden Konsequenzen diskutiert und auf die verschiedenen Auswirkungen auf die Verlässlichkeitsaspekte eingegangen. Der erste Teil schließt in einer Zusammenfassung der Ergebnisse in Thesenform. Der zweite Teil dieses Reports ist ein Glossar der in diesem Kontext wesentlichen Begriffe.

Wesentliche Erkenntnisse aus dem AP sind:

1. Systeme der künstlichen Intelligenz sind nicht ausschließlich Systeme des maschinellen Lernens, Systeme des maschinellen Lernens des sind nicht ausschließlich neuronale Netze.
2. Nichtlernende Systeme der künstlichen Intelligenz können bei der Nachweisführung analog den klassischen IT-Systemen der Anwendungsdomäne behandelt werden. Dies gilt weitgehend auch für Systeme des maschinellen Lernens, sowie die Lernmusterdaten eine klare Semantik besitzen.
3. Neuroinspirierte KI-Systeme unterscheiden sich von anderen komplexen IT-Systemen bezüglich der technischen Sicherheit nahezu ausschließlich durch die Rolle der Lernmusterdaten.
4. Der Lernvorgang bei Systemen des maschinellen Lernens ist als Teil der Softwareentwicklung zu betrachten und nimmt die Rolle der Modellübersetzung in der modellgetriebenen Softwareentwicklung oder des Codierens in der klassischen Programmierung ein.
5. Das größte Hindernis zur Überführung klassischer Ansätze und Verfahren zur Sicherstellung der Verlässlichkeit von Systemen mit neuronalen Netzen ist der Mangel an semantischen Modellen von Lernmusterdaten.
6. Derzeit können die geforderten Eigenschaften der Lernmusterdaten fast ausschließlich auf dem Weg von Audits und nur teilweise durch Testen verifiziert werden.

AP 4 Begleitung durch und Reflexion an bahnspezifischen Beispielen

In AP 4 sollten mittels Recherche von nationalen und internationalen Veröffentlichungen (Forschungsergebnisse, Konferenzbeiträge, Marktstudien) sowie zu Mindestanforderungen an die Funktionalität mögliche bahnspezifische Beispiele zum Einsatz von KI zusammengestellt und – nach Auswahl der konkret zu betrachteten Beispiele – offene Fragen bzgl. dieser gewählten Beispiele im Zulassungsprozess aufgezeigt werden.

Zunächst wurden einige mögliche Anwendungsbeispiele für KI-Systeme im Bahnsektor vorgestellt und diskutiert. Aus diesen die Beispiele wurden „Hinderniserkennung auf/neben dem Gleis per Kamera“ sowie „Bestimmung von Zugposition und -vollständigkeit per Fiber Optic Sensing“ ausgewählt.

Während das DB-Regelwerk bisher Zugfahrten ohne Triebfahrzeugführer nicht kennt, wird das automatisierte Fahren (ATO – automatic train operation) im Bahnbereich zunehmend thematisiert. In den höheren Automatisierungsgraden (GoA – Grade of Automation, vgl. IEC 62267) 3 und 4 ist kein Triebfahrzeugführer mehr an Bord. Das erste Beispiel konzentriert sich auf die Erkennung von Hindernissen auf oder neben dem Gleis mit Hilfe einer oder mehrerer auf dem Zug installierter Kameras. Diese zeichnen in regelmäßigen zeitlichen Abständen Bilder auf bzw. nehmen ein Video auf. Die Bilder bzw. das Video beinhalten einen Ausschnitt einer Bahnstrecke mit einem oder mehreren Gleisen und deren unmittelbarer Umgebung. Zur Erkennung und Klassifizierung von Hindernissen sowie der Bestimmung ihrer gleisgenauen Position, der Entfernung vom Zug und der Hindernisgröße[12] wird fortlaufend eine KI auf die Bilder angewandt, deren Ergebnisse jeweils spätestens einige Sekunden nach Aufzeichnung des ausgewerteten Bildes vorliegen. Aus Datenschutzgründen und aus Gründen des Datenvolumens erfolgt diese Auswertung ohne weiträumige Datenübertragung lokal am Standort der Kamera. Eine zeitlich beschränkte Historie von Bildern oder Ergebnissen kann in die aktuellen Ergebnisse einfließen, beispielsweise um Ergebnisse zu plausibilisieren oder Trajektorien bewegter Hindernisse zu bilden bzw. Bewegungsmuster zu erkennen. Die KI stellt ihre Ergebnisse inklusive Unsicherheiten einem lokalen oder entfernten System zur Verfügung, das, ggf. unter Berücksichtigung weiterer Informationen wie Zugpositionen und -geschwindigkeiten, eine angemessene Reaktion auslöst. Dies kann die Einleitung einer sofortigen Not- oder Betriebsbremsung von Zügen sein, eine Meldung an Züge in der Umgebung, die Benachrichtigung von Betriebs-, Wartungs-, Rettungspersonal oder Behörden, das Abwarten, ob sich das Ergebnis manifestiert, oder das Ignorieren des Ergebnisses. Ein entsprechendes System ist in Abbildung 3 skizziert.

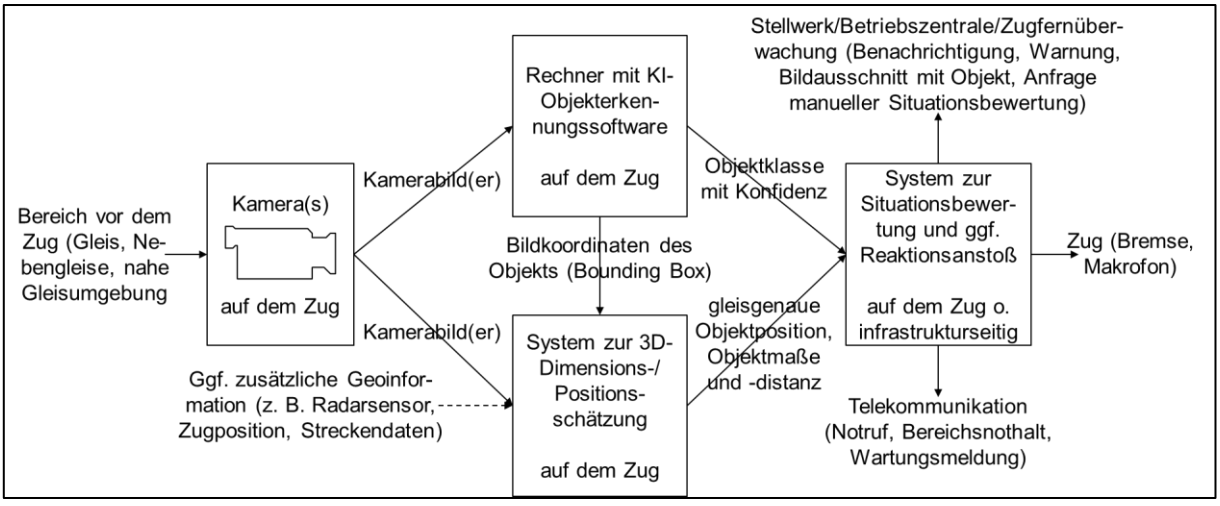


Abbildung 3: Skizze des Systems zur Hinderniserkennung

Je nach Betrachtungsfokus liegen alle vier rechteckigen Boxen (da sie gemeinsam den Triebfahrzeugführer ersetzen), lediglich die obere Box (da hier die KI zum Einsatz kommt) oder eine

[12] Die KI leistet klassischerweise die Objekterkennung (Identifikation von Bildregionen mit Objekten und Klassifizierung der Objektart). Zur Bestimmung der 3D-Position und -Größe des Objekts wird ein nachgelagertes (nicht-KI) System angenommen, das neben der durch die KI bestimmten Bounding Box weitere Informationen (z. B. Radar) als Eingabe haben kann.

Teilmenge der Boxen (wenn man z. B. die Fragestellung der Situationsbewertung und Reaktion als separates Problem betrachtet) innerhalb der Systemgrenzen.

Das Beispiel hat den Anspruch, die Streckenbeobachtung durch den Triebfahrzeugführer – ausgenommen die Beobachtung der bahntechnischen Anlagen – weitgehend zu ersetzen in dem Sinne, dass es möglicherweise für bestimmte Situationen einer Ergänzung durch andere technische Systeme bedarf (z. B. für Nachtsicht, für vereinzelte Ergänzung durch ortsfeste Beobachtung kritischer Punkte/Abschnitte an der Strecke oder für die Bereitstellung von Kontextinformationen zum Kamerabild). Solche anderen Systeme dürfen als vorhanden und ausreichend zuverlässig angenommen werden, soweit dies realistisch erscheint und die zugseitigen Kamerabilder zentrale Grundlage für die Hinderniserkennung bleiben.

Im anderen ausgewählten Beispiel geht es um die Positionsbestimmung und Vollständigkeitsprüfung durch so genanntes Fiber Optic Sensing. Es umfasst die regelmäßige Bestimmung (z. B. alle 2 Sekunden) der Zugpositionen und Zugvollständigkeiten aller auf einem zweigleisigen Streckenabschnitt von 20 km Länge verkehrenden Zügen per Fiber Optic Sensing. Die zeit- und ortsbezogenen Rohdaten werden dabei vorprozessiert, z. B. um die Amplituden näherer und entfernterer Signale zu normalisieren und die Signale zu glätten. Schließlich können aus der Intensität des Signals über dem Streckenabschnitt zum aktuellen Zeitpunkt t die Positionen der Zugmitten und damit die Zugpositionen bestimmt werden (auch ohne KI). Da die Signalqualität einer einzelnen Intensitätskurve über der Zeit (an einem festen Ort) oft nicht zur Prüfung der Zugvollständigkeit ausreicht und weil die Kurven für Orte, an denen sich der Zug aktuell befindet, noch nicht von allen Drehgestellen/Achsen überfahren wurden, werden dafür mehrere Kurven von hinter dem Zug liegenden Orten betrachtet. Die Kurven werden an der Ankunftszeit des ersten Drehgestells/Achsclusters aneinander ausgerichtet und eine Durchschnittsbildung durchgeführt. Für eine ausreichend genaue und einheitliche Bestimmung dieser Ankunftszeit wird eine KI eingesetzt. Schließlich wird die Anzahl der Peaks der resultierenden Kurve detektiert und mit der erwarteten bzw. vorherigen Anzahl der Drehgestelle/Achscluster des Zuges verglichen, um die Zugvollständigkeitsinformation zu erhalten. Ein Überblick über das System gibt Abbildung 4.

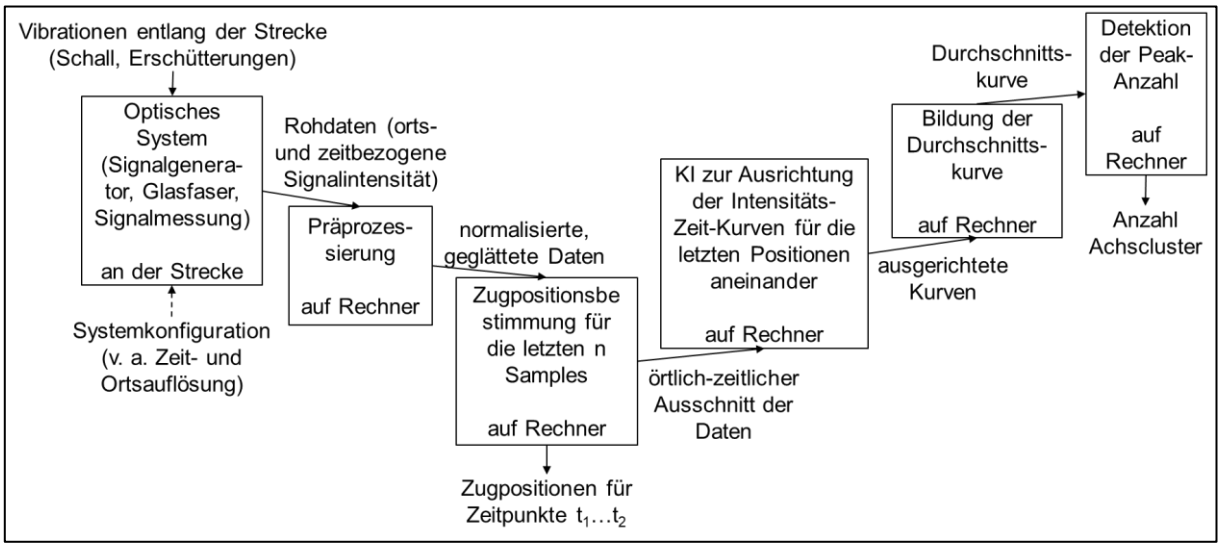


Abbildung 4: Skizze des Systems zur Bestimmung von Zugposition und -vollständigkeit[13]

[13] Wegen des hohen Datenvolumens soll davon ausgegangen werden, dass der/die Rechner in Nähe des Signalmessungssystems platziert und die Übertragung der Daten kabelgebunden stattfindet. Die Resultate (Zugpositionen, ermittelte Anzahl Achscluster) werden ebenfalls kabelgebunden an das übergeordnete Kontrollsystem (Stellwerk) weitergegeben.

Ziel ist es, auf dem Streckenabschnitt das Fahren im Moving Block auf Basis der erhaltenen Zugpositionen und Zugvollständigkeitsinformationen zu realisieren. Dabei wird von einem zentralen System ausgegangen, das Stellwerks- und ETCS Radio Block Center-Funktionalität integriert und basierend auf dem Status der Strecke Fahrerlaubnisse an Züge in seinem Bereich erteilt, die sich abzüglich eines Sicherheitspuffers bis zum (zuletzt gemeldeten) Ende des vorausfahrenden Zuges erstrecken können. Während die Meldung ETCS-konform per Positionsreport über Funk vorgesehen ist, wird sie im vorliegenden Fall durch das Fiber-Optic-Sensing-System übernommen. Gleisfreimeldeanlagen sind in Übereinstimmung mit lediglich für besonders kritische Abschnitte wie Weichen(-bereiche) bzw. zur Detektion des Eintritts/Austritts eines Zuges in/aus einen/einem Bereich vorgesehen.

Im Unterschied zur Objekterkennung ist allerdings die Korrelation zwischen den Parametern der Realität und denen der gemessenen Daten (Kamerabild bei Objekterkennung bzw. zeitlich-räumliches Signal bei FOS) weniger direkt; die Entfernung des Zuges von der Messeinrichtung, Achsabstände, Zuggeschwindigkeit, konkrete Quellen von Störsignalen wie Züge auf einem Parallelgleis etc. übersetzen sich in Signalstärke, Frequenz, Schwingungsmuster, Rauschen, Störsignal etc. Beide Ebenen haben ihre Vor- und Nachteile bei der Nutzung für die Formulierung von Anforderungen, und es ist wahrscheinlich, dass zur Erzeugung von Trainings- und Testdaten auch beide Ebenen genutzt werden (Kombination von eingemessenen und synthetisch erzeugten/kombinierten Signalen). Um am Ende jedoch dem Einsatz im realen Umfeld gerecht zu werden, scheint es zunächst einmal wichtig, auf der Ebene der Realität Anforderungen (an das Gesamtsystem inklusive KI) zu spezifizieren. Wichtige Parameter inklusive möglicher Grenzwerte finden sich in der obigen beispielhaften Umgebungsdefinition; die Granularität der Testwerte der meist kontinuierlichen Größen könnte durch eine in der Realität erwartete Verteilung sowie Randwerte definiert werden. Hinzu kommen Kalibrierungsparameter des FOS-Systems. Für letztere können Abhängigkeiten von anderen Parametern oder ein Kalibrierungsverfahren spezifiziert werden, um das System in verschiedenen Strecken- und Betriebskontexten nutzen zu können.

Auch bzgl. der Ausgabe unterscheidet sich das Beispiel des FOS von dem der Hinderniserkennung: Statt eines Vektors von Objektklassen mit Konfidenzen werden hier Skalenwerte durch die KI berechnet, und ob diese die „gewünschte“ (nicht näher definierte) Feinausrichtung der Samples aneinander ergeben, kann erst nach der Berechnung der Achsclusterzahl durch den Vergleich mit der tatsächlichen Achsclusterzahl des Zuges ermittelt werden. Verlässlich testen lässt sich die KI also erst als Teil des Gesamtsystems. Trotzdem lässt sich die Ausgabe der KI bereits anhand verschiedener Maße beurteilen (z. B. Vergleiche von aneinander ausgerichteter Samples auf minimale Differenzen ihrer Peakmitten oder auf maximalen Flächenüberlapp, oder Bewertung der Glätte der Mittelwertkurve mehrerer ausgerichteter Samples).

AP 5 Wechselwirkungen zwischen KI-Verfahren und Methoden zur Erprobung und Sicherheitsnachweisführung

In AP 5 sollten über Workshops unter den Projektbeteiligten zur Diskussion der Ergebnisse AP 1-3 anhand der Beispiele aus AP 4 eine Wissensbasis zum Einfluss verschiedener KI-Verfahren auf Test- und Zulassungsmethoden (als Basis für AP 6 und AP 7) erarbeitet sowie weiterer F&E-Bedarf geklärt werden.

In einer Vielzahl gemeinsamer Meeting (durch die Beschränkungen aufgrund der Covid-19-Pandemie lediglich als Online-Meetings) erfolgte eine Vorstellung und Diskussion der Ergebnisse vorgelagerter Arbeitspakete, um die Wechselwirkungen zwischen den KI-Verfahren, den jeweiligen

Zulassungssystemen sowie der Sicherheitsnachweisführung zu erörtern. Insbesondere Punkte von Relevanz für die Sicherheitsnachweisführung und Systementwicklung wurden hier diskutiert.

Im Rahmen der Diskussion der Systementwicklung nach DIN EN 50126 wurden nachfolgende Punkte besonders angesprochen:

1. Festlegung der Systemanforderungen
2. Entwurf und Implementierung
3. Systemvalidierung
4. Betrieb, Instandhaltung und Leistungsüberwachung

Im Rahmen der Diskussion der Sicherheitsnachweiserstellung lag der Fokus auf nachfolgenden Punkten:

1. Einleitung
2. Nachweis des korrekten funktionalen Verhaltens
3. Ausfallwirkungen
4. Betrieb mit externen Einflüssen
5. Ergebnisse der Sicherheitserprobung

Ergebnisse aus den Diskussionen sind in die weitere Bearbeitung der einzelnen Arbeitspakete eingeflossen.

AP 6 Mögliche Wege zur Nachweisführung von KI zum Erreichen der Sicherheitsanforderungen

In AP 6 sollten durch vergleichende Analysen mögliche Wege zur Nachweisführung von KI zum Erreichen der Sicherheitsanforderungen und die verbleibende offener Fragen (Lücken im Prozess) für fortführende Arbeiten beschrieben werden.

Im Rahmen dieses Paketes wurden Überlegungen angestellt, wie eine Systementwicklung nach DIN EN 50126 sowie die Sicherheitsnachweisführung nach DIN EN 50129 für sicherheitsrelevante KI-Anwendungen erfolgen könnte.

1. Systementwicklung nach DIN EN 50126

Phase „Systemkonzept“

In dieser Phase werden die Strategie und die Ziele der Systementwicklung festgelegt. Anwendungsbereich, Kontext und Zweck des Systems werden definiert. Ein wesentlicher Punkt ist die Analyse und Beschreibung der künftigen Systemumgebung, d. h. der technischen, physikalischen, geografischen und klimatischen Umgebungsbedingungen, der betrieblichen Randbedingungen, der Schnittstellen zu vorhandenen Systemen sowie der zu beachtenden Gesetze, Normen und Regelwerken. In dieser Phase erfolgt auch eine Analyse und Gegenüberstellung ähnlicher oder vergleichbarer Systeme. Dabei werden die Anforderungen an diese Systeme sowie deren technische und betriebliche Parameter sowie die Sicherheitsziele dieser Systeme ermittelt und auf ihre Anwendbarkeit bzw. Übertragbarkeit auf das neu zu entwickelnde System untersucht.

Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen. In AP 4 wurden beispielhaft Systemkonzepte bahnspezifischer Anwendungen skizziert.

Phase „Systemdefinition und betrieblicher Kontext“

Hier erfolgt eine Beschreibung des künftigen Systems einschließlich dessen Funktionalität und einer ersten Architekturskizze, aus der die Komponenten des Systems und die Verteilung der Aufgaben auf diese Komponenten hervorgehen. Es wird beschrieben, wie das System im Betrieb

angewendet werden soll. Hierzu gehören Betriebsverfahren, Betriebsarten, Use Cases zu verschiedenen betrieblichen Szenarien und die Instandhaltungsstrategie. Außerdem werden die klimatischen, mechanischen und elektrischen Einsatzbedingungen sowie die betrieblichen Umgebungsbedingungen spezifiziert. Es erfolgt eine Beschreibung der Systemgrenzen sowie der Schnittstellen und Wechselwirkungen zu anderen technischen Systemen und zum Menschen, d. h. dem Bediener, Instandhalter und Nutzer des Systems. Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen.

Phase „Risikoanalyse und -beurteilung“

In dieser Phase werden unerwünschte Ereignisse während des Betriebs des Systems und die daraus resultierenden Gefährdungen ermittelt. Hierzu werden zunächst alle Fehler identifiziert, die Ursache derartiger Ereignisse sein könnten, und die Fehlerfolgen hinsichtlich ihrer Kritikalität bewertet. Zu jeder Fehlerfolge wird das akzeptable Risiko ermittelt. Ist das tatsächliche Risiko höher als der akzeptierte Wert, werden Maßnahmen zur Risikoreduktion festgelegt, um sicherzustellen, dass das Risiko auf das akzeptable Maß reduziert wird.

Das akzeptierte Risiko wird als SIL (safety integrity level) und THR (tolerable hazard rate) angegeben. Der ermittelte SIL bestimmt den Umfang der Maßnahmen zur Vermeidung von systematischen Fehlern im Entwicklungsprozess (z. B. Dokumentation, Analysen, Tests, Verifikation und Validierung). Die THR legt die zulässige Rate gefährlicher Ausfälle fest und bestimmt somit die technische Gestaltung des Systems (z. B. Bauelementeauswahl, Redundanzprinzipien, Selbsttests, Überwachung im Betrieb).

Die Vorgehensweise zur Ermittlung des akzeptierten Risikos wird in den CSM-Verordnungen (EU) 402/2013 und (EU) 2015/1136 vorgeschrieben. Gängige Verfahren sind die Übernahme vorhandener Werte durch die Anwendung anerkannter Regeln der Technik oder der Vergleich mit einem bereits existierenden und akzeptierten Referenzsystem. Ist beides nicht möglich, muss das akzeptierte Risiko explizit ermittelt werden (z. B. über die Verfahren Risikograph oder Risikomatrix). Für technische Funktionen der Eisenbahnsicherungstechnik gilt die Vornorm DIN VDE V 0831-103 als anerkannte Regel der Technik. Dort werden die THR für typische Funktionen hergeleitet.

Für die Beispielanwendungen aus AP4 können die Anforderungen wie folgt hergeleitet werden:

Hinderniserkennung auf/neben dem Gleis per Kamera

- Vergleich mit der Funktion „Gefahrenraum freimelden“
- Die zu betrachtende Gefährdung ist: „Aufprall auf ein Hindernis im Gefahrenraum“.
- Die Bewertung der Barrieren erfolgt konservativ mit der Punktzahl 1.
- THR = 1E-06 h -1 (entspricht SIL1)

Im derzeit laufenden Forschungsprojekt „Risikoakzeptanzkriterien für den automatisierten Fahrbetrieb (ATO-RISK)“ werden die Risikoakzeptanzkriterien für den automatisierten Fahrbetrieb erforscht. Dazu zählt auch die Ermittlung des tolerablen Risikos und die Gewährleistung der mindestens gleichen Sicherheit, wenn man vom Fahren mit Triebfahrzeugführer auf automatischen Betrieb (ATO - Automatic Train Operation) wechselt. Nach Abschluss dieses Projekts können ggf. die dort ermittelten Werte übernommen werden.

Bestimmung von Zugposition und -vollständigkeit per Fiber Optic Sensing

Wird die KI nur zur Gleisfreimeldung verwendet und gibt es über der Gleisfreimeldung ein zusätzliches Sicherungssystem, welches die Ergebnisse plausibilisiert (z. B. Streckenblock), dann gilt:

- Vergleich mit der Funktion „Gleisabschnitt auf Freisein überwachen (Zugstraße)“
- Die zu betrachtende Gefährdung ist: „Zusammenstoß mit anderen Fahrzeugen“.

- Die Bewertung der Barrieren erfolgt mit der Punktzahl 2.
- THR = 1E-08 h -1 (entspricht SIL3)

Soll jedoch die vollständige Sicherung durch KI realisiert werden, dann gilt:

- Vergleich mit der Funktion „Schutz gegen Gegenfahrten sicherstellen (Zugfahrstraße)“
- Die zu betrachtende Gefährdung ist: „Zusammenstoß mit anderen Fahrzeugen“.
- Es gibt keine Barrieren.
- THR = 1E-09 h -1 (entspricht SIL4)

Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen. Mit der Festlegung eines SIL > 0 ergeben sich jedoch Anforderungen an die „sichere Softwareentwicklung“ der KI-Anwendung. Die Unterschiede zur Entwicklung einer „klassischen“ Software werden in den nachfolgenden Phasen betrachtet.

Phase „Festlegung der Systemanforderungen“

In dieser Phase werden die funktionalen und technischen Anforderungen an das zu entwickelnde System vollständig definiert. Hierzu gehören z. B.

a) Funktionen
b) Robustheit und Wartbarkeit
c) Leistung und Effizienz
d) Sicherheit
e) Schnittstellen
f) betriebliche Anwendung
g) zu berücksichtigende Umgebungsbedingungen (Klima, Mechanik, elektrische Bedingungen, EMV, physische und IT-Security Zugriffe)

Die Anforderungen müssen eindeutig, vollständig, widerspruchsfrei, korrekt, identifizierbar und prüfbar sein. Im Gegensatz zur klassischen Entwicklung mit präzisen, vollständigen Anforderungen soll KI dort eingesetzt werden, wo die Anforderungen nicht vollständig formuliert werden können (oder nur mit unverhältnismäßig hohem Aufwand). Die Anforderungsspezifikation der KI wird daher eher eine Beschreibung der zu beherrschenden Szenarien und dem jeweils erwarteten Ergebnis sein.

Offener Punkt für Nachfolgeprojekte

Gegenwärtig gibt es keine Kriterien, nach denen die Anforderungsspezifikation einer KI-Anwendung erstellt werden kann, so dass diese die Kriterien Vollständigkeit und Prüfbarkeit erfüllt.

Phase „Architektur und Aufteilung der Systemanforderungen“

In dieser Phase wird eine Systemarchitektur entwickelt, die die spezifizierten Anforderungen erfüllt. Die Anforderungen werden den Teilsystemen und Komponenten zugewiesen und es werden die Schnittstellen zwischen diesen Teilsystemen und Komponenten spezifiziert.

Die Architektur der in AP 4 beschriebenen Beispielapplikationen besteht im Wesentlichen aus

- Sensoren,
- einer Rechnerplattform inklusive Betriebssystem,
- der generischen KI-Software,
- der spezifischen KI-Software (d. h. den erlernten Daten und Parametern der generischen KI-Software) sowie
- der Ausgabeschnittstelle an die übergeordnete Anwendung.

Die Systemhardware inkl. der implementierten Redundanz- und Überwachungsmechanismen muss in der Lage sein, die geforderte THR zu erfüllen. Die Software (Betriebssystem, generische

und spezifische KI-Software) muss hinsichtlich enthaltener systematischer Fehler den geforderten SIL erfüllen. Hierzu sind bereits in der Architektur die Prinzipien zum Erreichen der geforderten Sicherheit festzulegen (z. B. fail-safe Hardware, Redundanz, Hard- und Software-Diversität).

Für COTS-Komponenten (z. B. Betriebssysteme oder Standardsoftware) werden in den Normen Kriterien definiert, nach denen deren Betriebsbewährung und Eignung für einen entsprechenden SIL beurteilt werden kann.

Offener Punkt für Nachfolgeprojekte

Die als generische KI-Software infrage kommenden Systeme sind gegenwärtig noch proprietäre Einzelanwendungen, die nicht die Kriterien der Betriebsbewährung und Eignung für COTS-Komponenten erfüllen. Um diese Software einzusetzen, ist eine umfassende Validierung erforderlich, in der nachgewiesen wird, dass die KI-Software frei von systematischen Fehlern ist oder die Wahrscheinlichkeit des Erkennens eines konkreten Sachverhalts begrenzt. Es gibt gegenwärtig kein Verfahren, mit dem eine hinreichende Validierung mit vertretbarem Aufwand möglich wäre.
Der Lernprozess zur Entwicklung der spezifischen KI-Software entspricht nicht den Anforderungen an die Softwareentwicklung gemäß EN 50128. Eine Möglichkeit zur Verringerung der Anforderungen an diesen Prozess wäre der Einsatz diversitärer KI-Software (z. B. Lernen mit diversitären Daten) und Vergleich der Ergebnisse oder eine zusätzliche Plausibilisierung der Ergebnisse der KI. Dies wurde jedoch in den vorangegangenen Arbeitspaketen nicht weiter untersucht.

Phase „Entwurf und Implementierung“

In dieser Phase erfolgt der Entwurf der Teilsysteme und Komponenten, so dass sie die an sie gestellten Anforderungen erfüllen. Die Hardwareentwicklung wird aus der weiteren Betrachtung ausgeklammert, da sie sich nicht von der Entwicklung konventioneller sicherer Systeme unterscheidet. Der Entwurf und die Implementierung der KI-Applikation entsprechen der „Lernphase“. Am Ende der Entwicklung muss diese Lernphase abgeschlossen sein, d. h., der Zustand der KI-Software wird „eingefroren“ und ein weiteres Lernen muss technisch ausgeschlossen werden.

Offener Punkt für Nachfolgeprojekte

Die klassische Softwareentwicklung nach EN 50128 wird bei einer KI-Anwendung durch den Lernprozess ersetzt. Folgende Fragen konnten in den vorangegangenen Arbeitspaketen noch nicht vollständig geklärt werden:

1) Wie kann die Aufgabenstellung an die KI-Anwendung auf einen vollständigen Satz prüfbarer Anforderungen und Lerndaten abgebildet werden?
2) Nach welchen Kriterien kann das Ende des Lernprozesses festgelegt werden, bzw. welche Anforderungen werden an Datenmenge, Datenqualität und Datendiversität gestellt, um sicher zu sein, dass die Aufgabenstellung am Ende des Lernprozesses vollständig und korrekt beherrscht wird?

Phase „Systemvalidierung“

Während der Systemvalidierung erfolgt die Prüfung und Bestätigung, dass das betrachtete System für den vorgesehenen Verwendungszweck geeignet ist und die festgelegten Anforderungen erfüllt. Die wesentlichen Mittel der Validierung sind Analysen und Tests.

Im Rahmen der Systemvalidierung sind zwei Sachverhalte nachzuweisen:

1) Hat die spezifische KI-Software die Anforderungen (Lerninhalte) korrekt im Sinne der Aufgabenstellung interpretiert? Hierzu ist ein analytischer Nachweis erforderlich.
2) Werden alle Anforderungen an die KI-Software vollständig und korrekt erfüllt? Dieser Nachweis erfolgt durch entsprechende Tests.

Offene Punkte für Nachfolgeprojekte

Die folgenden Fragen konnten in den vorangegangenen Arbeitspaketen nicht geklärt werden und bieten sich für eine weitergehende Betrachtung an:

1) Wie und in welchem Umfang kann analytisch nachgewiesen werden, „was“ die KI-Software gelernt hat?
2) Nach welchen Kriterien kann ein hinreichender Testumfang definiert werden, um eine vollständige Anforderungsüberdeckung durch Tests zu erreichen?
3) In welcher Testumgebung können bzw. müssen die Tests ausgeführt werden (Labor, Simulation, Feld)?
4) Wie kann eine hinreichende Unabhängigkeit zwischen „Lerndaten“ und „Testdaten“ erreicht werden, um systematische Fehler aufgrund gleicher und ggf. unzureichender Datenquellen zu vermeiden?

Phase „Systemabnahme“

In dieser Phase erfolgt die abschließende Begutachtung bzw. Bewertung des Systems und dessen Abnahme durch den Betreiber. Grundlage hierfür bilden die Nachweise der vorangegangenen Phasen und ggf. ergänzende Prüfungen oder Tests. Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen

Phase „Betrieb, Instandhaltung und Leistungsüberwachung“

Zu dieser Phase gehören: Betrieb des Systems, Schulung des Personals, Instandhaltung des Systems, Fehlermanagement (Analyse, Bewertung, Korrektur) sowie Führen des Gefährdungslogbuchs.

Offene Punkte für Nachfolgeprojekte

Es müssen entsprechende Verfahren (Anleitungen) für Betrieb und Instandhaltung der generischen und spezifischen KI-Software entwickelt werden. Hierzu sind folgende Fragen zu klären:

1) Wodurch unterscheiden sich die Verfahren zum Betrieb und der Instandhaltung eines KI-Systems von denen eines konventionellen Systems?
2) Wie kann der Fehlermanagementprozess eines KI-Systems definiert werden (d. h. das Verfahren zum Erkennen von Fehlern der KI-Software, zur Analyse der Fehlerursache und der Fehlerauswirkungen)?
3) Welche Maßnahmen sind zur Fehlerbehebung der KI-Software erforderlich (z. B. weiteres Lernen oder „Löschen“ oder „Korrigieren“ fehlerhaften Wissens)?

Phase „Außerbetriebnahme“

In der Phase Außerbetriebnahme erfolgen das Abschalten des Systems am Ende seiner Nutzung und das Entfernen aus seiner Systemumgebung. Die Anforderungen an diese Phase gelten zunächst gleichermaßen für konventionelle Systeme und KI-Anwendungen. Die KI-Software ist in ihren Ausgangszustand vor Beginn der Lernphase zu versetzen.

2. Sicherheitsnachweisführung nach DIN EN 50129

Abschnitt „Einleitung“

In diesem Abschnitt erfolgt ein Überblick über das System und den Systementwurf und eine Darstellung der Prinzipien, auf denen sich die Sicherheit des Systems abstützt. Die Inhalte leiten sich aus den Ergebnissen der Phasen „Systemkonzept“ bis „Festlegung der Systemanforderungen“ ab.

Offene Punkte für Nachfolgeprojekte

Die Voraussetzung für die Sicherheitsnachweisführung ist neben einer hinreichenden Systembeschreibung eine Klärung der auf KI anwendbaren Sicherheitsprinzipien. In den vorangegangenen

Arbeitspaketen wurden keine speziellen Sicherheitsprinzipien für KI-Software identifiziert. Nach Abschluss der Lernphase befindet sich die KI-Anwendung in einem „eingefrorenen“ Zustand. Ihr Verhalten wird nicht von stochastischen Größen beeinflusst, so dass sie zu jedem Zeitpunkt mit identischen Eingangsdaten auch identische Ergebnisse liefert. Damit unterscheidet sie nichts von einer herkömmlichen Software-Anwendung. Der Unterschied zur klassisch entwickelten Software besteht darin, dass der Lernprozess nicht mit den Methoden der klassischen Softwareentwicklung verifiziert werden kann. Es konnte noch nicht geklärt werden, wie die Qualität des Lernprozesses quantifiziert werden kann.
Ein möglicher Lösungsweg, auch mit nicht vollständig verifizierter Software eine hinreichende Sicherheit zu erzielen, ist der Einsatz diversitärer Software. Hierzu müssten diversitäre KI-Anwendungen mit unterschiedlichen Lernansätzen eingesetzt oder diese mit unterschiedlichen Arten von Lerndaten trainiert werden. Es konnte nicht geklärt werden, inwieweit dies möglich ist und in welchem Umfang bei dieser Lösung Common-Cause Effekte berücksichtigt werden müssen.

Abschnitt „Nachweis des korrekten funktionalen Verhaltens“

Der wesentliche Inhalt dieses Abschnitts ist der Nachweis zur Erfüllung der funktionalen Anteile der System-Anforderungsspezifikation (d. h. der funktionalen betrieblichen Anforderungen) und der Sicherheits-Anforderungsspezifikation (d. h. der funktionalen Sicherheitsanforderungen). Hierzu gehören der Nachweis der korrekten Hardware-Funktionalität und der Nachweis der korrekten Software-Funktionalität. Letzterer ist für die KI-Anwendung relevant. Die anzuwendende Norm ist die DIN EN 50128. Der Nachweis der korrekten Softwarefunktionalität erfolgt durch Analysen und Tests.

Angewendet auf eine KI-Anwendung bedeutet dies, dass in einem analytischen Nachweis zu zeigen ist, dass die KI-Software „das Richtige“ gelernt hat. Die Ergebnisse dieses analytischen Nachweises sind durch Tests zu plausibilisieren. An die Tests werden folgende Anforderungen gestellt:

- vollständige Anforderungsüberdeckung
- Unabhängigkeit der Testentwicklung von der Systementwicklung

Offene Punkte für Nachfolgeprojekte

Der analytische Nachweis erfolgt bei herkömmlicher Software über eine durchgehende Anforderungsverfolgung von der Spezifikation über das Design bis zum Code. Dieser Weg ist bei einer KI-Anwendung nicht möglich. Hierzu muss z. B. geklärt werden, wie aus den Parameterbelegungen der KI-Software am Ende der Lernphase auf die erlernten Inhalte geschlossen werden kann.
Eine weitere offene Frage ist, wie auf Basis einer typischerweise „unscharfen“ Anforderungsspezifikation der KI-Anwendung ein Nachweis der vollständigen Anforderungsüberdeckung geführt werden kann. Um die Unabhängigkeit der Testentwicklung von der Systementwicklung zu erreichen, ist es notwendig, die Unabhängigkeit des Lernprozesses und der Lerndaten vom Testprozess mit Testumgebung und Testdaten nachzuweisen. Alle verbleibenden Unschärfen oder Fehler der KI-Software sind systematische Fehler. Sie beeinflussen direkt den erreichbaren SIL. In der Norm EN 50129 wird gefordert, dass Fehlzustände aufgrund systematischer Fehler erkannt und zu einer sicherheitsgerichteten Ausfallreaktion führen müssen. Dies kann z. B. durch eine Plausibilisierung der Ergebnisse durch eine zusätzliche Überwachungseinrichtung erfolgen. Dies wird nicht in allen Anwendungsfällen möglich sein und schränkt die Einsetzbarkeit von KI bei hohen Sicherheitsanforderungen (SIL) entsprechend ein. Die generische KI-Software wird im Sicherheitsnachweis als „pre-existing Software“ behandelt. Hierzu muss es gewisse Qualitätsanforderungen erfüllen. Gegenwärtig gibt es keine professionelle KI-Software, mit der ein solcher Nachweis durchführbar ist.

Abschnitt „Ausfallauswirkungen“

In diesem Abschnitt ist insbesondere die Ungefährlichkeit von Hardwareausfällen nachzuweisen. Für KI-Applikationen ist der Teil „Schutz vor systematischen Fehlern“ relevant. Wenn davon ausgegangen werden muss, dass die KI-Software aufgrund der Unmöglichkeit des vollständigen Nachweises des korrekten funktionalen Verhaltens systematische Fehler enthält, dann ist hier zu zeigen, wie solche systematischen Fehler erkannt und in einen sicheren Zustand überführt werden. Eine Möglichkeit hierzu ist eine zusätzliche Überwachungseinrichtung, welche Fehler in den Ergebnissen der KI erkennt und daraufhin eine sichere Reaktion einleitet.

Wird zum Schutz vor gleichgerichteten systematischen Fehlern in der KI-Software der Ansatz diversitärer KI-Software verfolgt, dann ist hier die Unabhängigkeit der diversitären KI zu zeigen. Dies betrifft insbesondere den Schutz vor gleichgerichteten systematischen Fehlern. Dieser kann z .B. erreicht werden durch eine Kombination von:

- diversitärer generischer KI-Software
- diversitären Lernprozessen
- diversitären Lerndaten

Offene Punkte für Nachfolgeprojekte

Es ist gegenwärtig nicht geklärt, wie eine hinreichende Diversität (Unabhängigkeit) der verschiedenen verfügbaren generischen KI-Software nachgewiesen werden kann. Gleiches gilt für die Entwicklung diversitärer Lernprozesse und Lerndaten.

Abschnitt „Betrieb mit externen Einflüssen“

In diesem Abschnitt erfolgt der Nachweis, dass das entwickelte System auch unter Einwirkung der anzunehmenden externen Einflüsse seine betrieblichen Anforderungen und seine Sicherheitsanforderungen erfüllt. Externe Einflüsse sind z. B. Umgebungsbedingungen wie Temperatur, Luftfeuchte, mechanische und elektrische Einflüsse. Je nach Art der verwendeten Sensoren kommen ggf. auch optische oder akustische Einflüsse hinzu. Der Nachweis, dass die KI-Software unabhängig von den Umgebungsbedingungen korrekte Ergebnisse liefert, erfolgt überwiegend durch Tests.

Offene Punkte für Nachfolgeprojekte

Ein erhoffter Vorteil der Anwendung von KI ist das Identifizieren charakteristischer Merkmale der Eingangsgrößen trotz nicht vollständig beschreibbarer Umgebungsbedingungen. Auch hier stellt sich daher die Frage, wie ein hinreichender Lern- und Testdatensatz definiert werden kann, um eine vollständige Überdeckung der zu erwartenden Umgebungsbedingungen zu erreichen. Dieser Testdatensatz muss wieder ebenfalls „unabhängig“ von den Lerndaten entwickelt werden.

Abschnitt „Sicherheitsbezogene Anwendungsbedingungen“

In diesem Abschnitt erfolgt die Definition aller Regeln, Bedingungen und Einschränkungen, die bei der Anwendung des Systems zu beachten sind, damit dessen Sicherheit gewährleistet ist. Insbesondere müssen hier die aus den Restriktionen der KI-Anwendung resultierenden Einschränkungen und Anwendungsregeln hergeleitet und definiert werden.

Offene Punkte für Nachfolgeprojekte

Es ist anzunehmen, dass KI gegenwärtig noch nicht in der Lage ist, die geforderten Funktions- und Sicherheitsziele vollständig zu erreichen. Bezogen auf die Anwendungsbeispiele der vorangegangenen Arbeitspakete konnte noch nicht geklärt werden, welche Restriktionen der KI hier zu erwarten sind und durch welche Maßnahmen diese kompensiert werden können.

Abschnitt „Ergebnisse der Sicherheitserprobung“

Im Anschluss an den theoretischen Nachweis durch Analysen und Tests erfolgt die Sicherheitserprobung. Diese dient dazu, das Vertrauen in das neu entwickelte System zu stärken. Ihr Ziel ist eine hinreichende Erprobung des Systems unter allen relevanten Betriebsbedingungen. Zu beachten ist, dass die Sicherheitserprobung nicht mit den Tests zum Nachweis des korrekten funktionalen Verhaltens und des Betriebs mit externen Einflüssen gleichzusetzen ist. Diese Tests müssen vor Beginn der Erprobung erfolgreich abgeschlossen sein. Es genügt daher nicht, die KI im Feld zu „erproben“ und darauf dann die Sicherheit zu begründen.

Man unterscheidet eine Sicherheitserprobung mit und ohne Sicherheitsverantwortung des zu erprobenden Systems. Bezogen auf die KI-Anwendung ist es daher möglich, das System parallel zum bereits existierenden System zu erproben und während dieser Phase die Ergebnisse des KI-Systems mit denen des bestehenden Systems zu vergleichen.

Offene Punkte für Nachfolgeprojekte

Es wurde noch nicht untersucht, über welchen Zeitraum und unter welchen Bedingungen eine Sicherheitserprobung durchgeführt werden muss, damit am Ende der Erprobung ein hinreichendes Vertrauen in das System vorliegt. Dies steht insbesondere im Zusammenhang mit der noch ungeklärten Frage einer hinreichenden Testabdeckung während der Systemvalidierung.

3. Resümee

Die Untersuchungen im Rahmen des Projektes haben gezeigt, dass es möglich ist, die Entwicklung einer KI-Entwicklung auf den Entwicklungsprozess gemäß EN 50126, EN 50128 und EN 50129 abzubilden. Es sind jedoch noch Fragen bzw. offene Punkte vorhanden, die im Rahmen von Nachfolgeprojekten zu klären sind. Diese betreffen insbesondere:

- Wie kann eine Anforderungsspezifikation für KI-Anwendungen so erstellt werden, dass diese das Kriterium „Vollständigkeit“ erfüllt?
- Wie kann ein analytischer Nachweis der Anforderungserfüllung erfolgen, d. h., die Frage beantwortet werden „WAS“ die KI gelernt hat und ob damit alle Anforderungen erfüllt werden?
- Wie kann ein hinreichender Satz an Lern- und Testdaten zum vollständigen Nachweis der Anforderungsüberdeckung erstellt werden?
- Wie kann die Unabhängigkeit von Lern- und Testdaten gewährleistet und nachgewiesen werden?

Mögliche Lösungsansätze zur Beherrschung sind die Entwicklung diversitärer KI-Anwendungen und die Überwachung der KI-Anwendung durch eine zusätzliche Überwachungseinrichtung. Für den Einsatz zweier diversitärer KI-Anwendungen sind folgende Fragen bzw. offene Punkte zu klären:

- Wie kann die Unabhängigkeit der verwendeten generischen KI-Software nachgewiesen werden?
- Wie kann die Unabhängigkeit der Lernprozesse beider KI-Anwendungen nachgewiesen werden?
- Wie kann die Unabhängigkeit der Lerndaten beider KI-Anwendungen nachgewiesen werden?
- Verbleiben trotz des diversitären Ansatzes noch Common-Cause Effekte und wie können diese identifiziert werden?

AP 7 Ansätze zur Überarbeitung bzw. Anpassung der Zulassungsvorschriften

In AP 7 sollten mittels Gesetzes- und Normanalyse Änderungsbedarfe in den Verordnungen, Gesetzen, TSI bzw. Produktnormen zum Einsatz von KI in sicherheitskritischen Anwendungen festgestellt werden.

Im Zusammenspiel mit AP 6 wurden die spezifischen Normen und Gesetze betrachtet und ein Konkretisierungsbedarf hinsichtlich der Normen festgestellt, insbesondere zur Systementwicklung nach DIN EN 50126 sowie zur Sicherheitsnachweisführung nach DIN EN 50129 für sicherheitsrelevante KI-Anwendungen. Vorschläge, in welche Richtung die einschlägigen Normen weiterentwickelt werden könnten, wurden unterbreitet. Ferner wurden Defizite im Zusammenspiel zwischen Normung und bahnspezifischer Regulierung aufgedeckt. Insbesondere seitens der europäischen Gesetzgebung besteht erheblicher Verbesserungsbedarf seitens der Formulierungen und Übersetzungen der Normen. Diese sind teilweise inkohärent gefasst und uneinheitlich übersetzt, was Missverständnisse befördert. Ferner sollte im Rahmen der Mobilitätssektoren ein Gleichlauf angestrebt werden, insbesondere weil in den Mobilitätskonzepten der Zukunft eine engere Verzahnung dieser verschiedenen Mobiltätssektoren stattfinden wird und sich die KI-bezogenen Anwendungsszenarien ähneln (z. B. bei der Objekterkennung).

2 Wichtigste Positionen des zahlenmäßigen Nachweises

Das Forschungsvorhaben wurde an der TU Chemnitz durch die Professuren Medieninformatik (Prof. Dr. Maximilian Eibl), Betriebssysteme (Prof. Dr. Matthias Werner) sowie Privatrecht und Recht des geistigen Eigentums (Prof. Dr. Dagmar Gesmann-Nuissl) und dem Zentrum für Wissens- und Technologietransfer der TU Chemnitz (ZWT) durchgeführt. Über eine Auftragsvergabe waren das Institut für Verkehrsforschung des Deutschen Zentrums für Luft- und Raumfahrt e. V. (DLR), die Ingenieurgesellschaft für Verkehrssicherungstechnik GmbH (IVS), das Fraunhofer-Institut für Werkzeugmaschinen und Umformtechnik (IWU), das Fraunhofer-Institut für Elektronische Nanosysteme (ENAS), die Siemens Mobility GmbH, die IFB Institut für Bahntechnik GmbH, die IAV GmbH Ingenieurgesellschaft Auto und Verkehr sowie der Smart Rail Connectivity Campus e. V. eingebunden.

Die im Rahmen des Projektes entstandenen Ausgaben betrugen insgesamt 252.925,28 € (zzgl. Projektpauschale i. H. v. 50.585,06 €). Sie liegen damit 17.563,34 € (ohne Projektpauschale) unter der bewilligten Zuwendung. Die Ausgaben wurden im Wesentlichen für Personalausgaben zur Beschäftigung wissenschaftlicher Mitarbeiter an der TU Chemnitz (121.205,17 €) sowie für die Auftragsvergabe zur Einbindung externer Kompetenzträger (131.424,73 €) verwendet.

Dies entsprach auch in etwa der ursprünglichen Finanzplanung:
Geplant waren Personalausgaben i. H. v. 130.479,88 €, so dass 9.274,71 € weniger benötigt wurden. Dies ist insbesondere auf eine z. T. erst verspätet mögliche Besetzung von Personalstellen zurückzuführen, die im Laufe des Projektes nicht ausgeglichen werden konnte. Aufgrund von Vorarbeiten bzw. verstärkte Einbindung externer Kompetenzträger konnten die geplanten Projektergebnisse dennoch erreicht werden.
Zur Einbindung externer Kompetenzträger standen 134.008,74 € zur Verfügung, von denen 2.584,01 € nicht verausgabt worden. Aufgrund der mit der Covid19-Pandemie einhergehenden Einschränkungen wurden die Meetings ausschließlich online durchgeführt, so dass die Auftragnehmer keine Reisekosten aufbringen mussten. Gleichzeitig wurden die Kompetenzträger aber z. T. intensiver als geplant eingebunden, so dass dieser Mehraufwand die geplanten Auftragswerte wieder nahezu ausgeglichen hat.

Ausgaben für Dienstreisen im Inland wurden mit 295,38 € wesentlich weniger benötigt als geplant (6.000 €). Dies ist auf die mit der Covid19-Pandemie einhergehenden (Reise-)Einschränkungen zurückzuführen.

3 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Die durchgeführten Forschungsarbeiten im Projekt SRCC-KI und die dafür aufgewendeten Ressourcen waren notwendig und angemessen, da sie der Planung laut Projektantrag entsprachen und alle wesentlichen im Arbeitsplan formulierten Aufgaben erfolgreich bearbeitet wurden. Mittel wurden insbesondere bei den geplanten Personalausgaben und den Ausgaben für Dienstreisen im Inland eingespart. Es waren keine zusätzlichen Ressourcen für das Projekt notwendig.

4 Voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse

Das Befähigerprojekt „KI-bezogene Test- und Zulassungsmethoden“ trägt maßgeblich dazu bei, dass der Smart Rail Connectivity Campus langfristig nachhaltig etabliert werden und damit einen Beitrag zum Strukturwandel in der Region Chemnitz-Erzgebirge leisten kann. Um dies zu erreichen, wurden mit dem vorliegenden Vorhaben Grundlagen in Form einer breiten Wissensbasis erarbeitet und konkrete F&E-Bedarfe abgeleitet, die dann im Zuge weiterer F&E-Vorhaben konkretisiert und implementiert werden müssen. Eine direkte wirtschaftliche Verwertung der vorliegenden Ergebnisse ist nicht möglich (und war auch nicht angedacht).

Zudem wird mit dem Anwenderkompass eine Hilfestellung für Unternehmen gegeben, die sich bisher noch nicht oder nur selten mit Zulassungsprozessen für bahntechnische Innovationen beschäftigt haben – dies trägt direkt zu einer Förderung des Wissens- und Technologietransfers in die (regionale) Wirtschaft und zur Anwendung der Forschungsergebnisse bei. Dadurch werden die F&E-Kompetenzen und -kapazitäten insbesondere auch der Unternehmen in der Region gestärkt und die Basis für die bessere Ausnutzung vorhandener Wertschöpfungspotenziale gelegt.

5 Erfolgte oder geplante Veröffentlichungen

Die Professur „Privatrecht und Recht des geistigen Eigentums“ der TU Chemnitz hat einen Beitrag in der Zeitschrift Digital Society mit dem Titel „Auditing of AI: A European Legal Approach“ veröffentlicht. Zudem plant die Professur eine Veröffentlichung Ende 2022 in der InTeR-Schriftenreihe (Zeitschrift zum Innovations- und Technikrecht), in welche die Projektergebnisse mit einfließen werden.

Sachbericht zum Verwendungsnachweis

Anhänge

WIR!-Projekt:	**„KI-bezogene Test- und Zulassungsmethoden (SRCC-KI)” (FKZ 03WIR1205)**
Laufzeit:	**08/2020 – 01/2022**

AP1 - KI-Praxisbeispiele, Erfahrungen anderer Branchen und Schutzrechtsanalyse

Josef Haupt

Technische Universität Chemnitz, Medieninformatik

1 Einleitung

Im Rahmen der Bearbeitung von AP 1 wurde Rechere zu möglichen unerwünschten Verhaltensweisen, mit dem Schwerpunkt auf sog. Adversarial Attacks (AA), von Neuronalen Netzen (NN) betrieben. Im Rahmen der Literaturrecherche wurde ein Überblick über die Gefahren und Angriffsvektoren, welche durch moderne technische Systeme, entstehen. Aktuelle Angriffsstrategien erlauben dem Angreifer auch das manipulieren von geschlossenen Systemen (Blackbox). In dieser Zusammenfassung der Angriffstechniken, werden auch Gegenmaßnahmen vorgestellt, welche teilweise in der Lage sind Angriffe abzuwehren.

In Zusammenarbeit mit den Projektpartnern wurde erkannt, dass eine Verifikationen eines neuronalen Netzes nur möglich ist, wenn diesem bestimmte Eigenschaften nachgewiesen werden können. In dem Sinne wurde ebenfalls eine Literaturrecherche zum Thema Nachweisbarkeit von tiefen neuronalen Netzen vorgenommen. Diese zeigte, dass bereits Forschungsgebiete existieren deren Ziel es ist, das Verhalten von künstlicher Intelligenz (KI) für menschliches Verstehen aufzubereiten. Der Bereich der "Explainable AI" versucht die Ausgaben von KI zu interpretieren bzw. zu zeigen wieso ein Ergebnis zustande kam. Im Arbeitspaket wird eine Überblick der Möglichkeiten vorgestellt, die in der aktuellen Literatur zu finden waren.
Die recherchierten Themen und Techniken sind nicht im speziellen für den Einsatz im Bereich des autonomen Fahrens entwickelt, sondern sind lediglich eine Erweiterung der Probleme, welche entstehen, wenn neuronale Netze in komplexe Systeme eingeführt werden, welche externen Eingaben erlauben.
Neben den technischen Aspekten zum Thema KI, enthält diese Zusammenfassung auch einen Überblick zum Aufbau von automonen Fahrsystemen, wie sie in der aktuellen Literatur beschrieben werden.

2 Technische Verfahren aus dem Bereich des autonomen Fahrens

Um autonomes Fahren zu ermöglichen werden Systeme benötigt, welche in der Lage sind, basierend auf den Daten der Sensoren, Instruktionen für das Fahrzeug zu generieren. Generell werden diese Computing Systems (CS) in zwei Gruppen unterteilt [5].

Modulbasierte Verfahren Das CS wird in separate Module unterteilt, dabei erfüllt jedes Modul eine gesonderte Funktion. Dies ermöglicht das Zusammenarbeiten von Menschen mit unterschiedlichen Schwerpunkten [15], sowie eine Abgrenzung der Funktionsbereiche.
Generell unterscheiden sich solche Systeme hauptsächlich anhand ihrer Software und der Konfiguration der Sensoren. Zu bekannten Vertretern von modulbasierten Verfahren gehören u. a. Boss [26], Google [2], TerraMax und BRAiVE [5].

Ende-zu-Ende Lernverfahren Die Pipeline für Ende-zu-Ende Verfahren ist weitaus einfacher. Hierfür werden die Sensordaten direkt an eine entsprechend vorher trainierte künstliche Intelligenz übergeben. Diese generiert anhand der Daten direkt Instruktionen zur Steuerung des Fahrzeuges.
Diese Verfahren wurden vor allem durch den rasanten Fortschritt im Bereich von maschinellem Lernen (ML) angestoßen. Trotz der vermeintlich einfacheren Architektur des Autonomous Driving System (ADS) durch Ende-zu-Ende Verfahren, werden diese weitestgehend nicht in aktuellen Anwendungen verwendet [29]. Ende-zu-Ende-Verfahren sind dennoch Gegenstand aktueller Forschung (siehe Tabelle 1) und könnten in Zukunft, aufgrund ihrer Einfachheit mehr Einsatz finden.

Quelle	Methode
[18]	Mobiler Roboter lernt basierend auf Bilddaten und Bewegungswinkel Hindernissen auszuweichen.
[3]	Ein neuronales Netz überführt die Bilddaten der Kamera direkt zu Kommandos.
[21]	Automatisiertes Fahren in einer simulierten Umgebung.

Table 1. Die Tabelle enthält Beispiele bei denen Ende-zu-Ende Lernverfahren angewendet wurden.

2.1 Architektur eines Autonomous Driving System (ADS)

Ein ADS besteht generell aus vier Teilen. Den Sensoren, dem Betriebssystem (OS), der Middleware und den Anwendungen des Systems. Diese einzelnen Kom-

ponenten haben unterschiedliche Aufgaben, auf welche im Weiteren eingegangen wird.

Sensoren Die Sensoren sind die "Augen und Ohren" des Systems und der Start der ADS-Pipeline. Um ein möglichst umfassendes und vollständiges Abbild der Umgebung des Fahrzeuges erzeugen zu können, verfügen autonome Fahrzeuge über eine Reihe von Sensoren mit unterschiedlichen Schwerpunkten und Funktionsweisen, siehe Tabelle 2.

Sensor	Beschreibung
Kamera	Sehr beliebt, aufgrund des geringen Preises und der hohen Nutzbarkeit.
Radar	Teilweise teurer als eine Kamera, allerdings auch weniger von Wetter und Belichtung abhängig.
LiDAR	Misst, genau wie das Radar, die Distanz zur Umgebung, mit dem Unterschied, dass hier ein Laser verwendet wird, statt elektromagnetischer Wellen.
GPS/GNSS	Dient vor allem der Lokalisierung des Fahrzeugs.

Table 2. Die Tabelle enthält Sensortechnik, die in modernen ADS verbaut ist.

Betriebssystem Über Treiber werden im OS Brücken zwischen Hard- und Software erstellt. In sicherheitskritischen Szenarios, wie dem autonomen Fahren, muss das Betriebssystem Echtzeit-Anforderungen erfüllen.

Middleware Während das OS die physischen Komponenten mit der Software verbindet, regelt die Middleware die Kommunikation zwischen dem Betriebssystem und den Anwendungen. Es werden dabei Schnittstellen zur Verfügung gestellt die zur Programmierbarkeit zum System beisteuern, dadurch wird es möglich das System effizienter zu entwickeln und zu erweitern.

Anwendungen Die tatsächliche Erfassung und Steuerung erfolgt über die Anwendungsschicht, welche auf OS und Middleware aufsetzt. Die typische Anwendungen eines ADS sind z.B. Spurhaltung, -erkennung, Vorhersage, Planung und Fahrzeugsteuerung. Über die Anwendungen werden Steuer-Kommandos generiert und zum Fahrzeug weitergeleitet.

3 Überblick aktueller Stand der Technik

3.1 Nachweisbarkeit von neuronalen Netzen

Tiefe neuronale Netze (DNN) sind durch nicht-lineare Schichten und ihre hohe Anzahl von miteinander verknüpften Neuronen in der Lage komplexe Funktionen zu lernen und abzubilden. Typischerweise werden sie daher an Bereichen angewendet in denen Menschen nicht in der Lage sind, einen allgemeines Regelsatz für die Daten zu konstruieren und eine Vielzahl von Beispieldaten zum Trainieren des DNN verfügbar sind.
Die nicht-lineare Natur von neuronalen Netzen hat den Nachteil, dass ein direkter Nachweis eines DNN oft nicht möglich ist. Meistens ist es ausreichend das DNN als Black-Box zu behandeln. In Anwendungen in denen die Sicherheit ein entscheidender Faktor ist, kann allerdings vom menschlichen Betrachter nicht erkannt werden, welche Teile der Eingabe die Entscheidung der KI beeinflusst haben. Das Forschungsgebiet der *Explainable AI* beschäftigt sich mit Techniken, welche die Entscheidungen und Ausgaben von DNN für Menschen erklärbar machen soll. Viele Veröffentlichungen setzen dabei auf Visualisierung oder die Berechnung von Scores, anhand welcher ein menschlicher Beobachter erkennen kann was genau das Modell beeinflusst um entsprechende Fehlentscheidungen zu erklären.
Die Methoden können anhand des Zeitpunktes der Erklärung in zwei Bereiche eingeteilt werden.

posthoc Die Interpretation basiert auf der Ausgabe des Netzwerkes.

ante-hoc Die Struktur der künstlichen Intelligenz produziert bereits eine Erklärung oder ist von Natur aus erklärbar.

Visualisierung Methoden der Visualisierung versuchen die Bereiche der Eingabe in den Fokus zu bringen, welche das DNN am stärksten beeinflussen. Dadurch können für das Netz relevante Teil der Daten erkannt werden. Anhand von Visualisierung können somit zum Beispiel Fehler in der Feature-Extraktion innerhalb des Netzes erkannt werden. Die Daten könnten für Menschen unerkennbare Zusammenhänge haben, welche das Netz verwendet um seine Fehlerquoten auf den Trainingsdaten zu verringern, die aber für das Grundproblem irrelevant sind. So lernt z.B. eine KI Blumen anhand des Hintergrunds im Datenset zu unterscheiden, was jedoch für die Anwendung in der realen Welt oft nicht erwünscht ist.

Gradients Die *Gradients*-Methode ist ein posthoc Ansatz zu Visualisierung [22]. Hierfür wird der Gradient der Vorhersage im Bezug auf die Eingabe mit festen Gewichten berechnet. Der Hauptanwendungsfall dieser Technik liegt in der Bildverarbeitung (siehe Fig. 1).

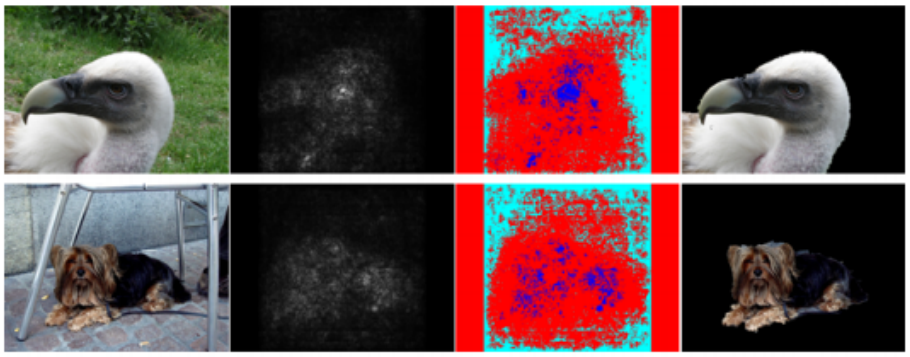


Fig. 1. Die Gradients-Methode zeigt über die Gradienten des DNN bezüglich eines Samples, welche Gebiete viel Einfluss auf die Ausgabe innerhalb der Schichten haben.

Guided Backpropagation Die *Guided Backpropagation* ist ein posthoc Ansatz [23]. Hierfür wird die Gradients-Technik mit DeconvNets kombiniert, was die für das Netz markanten Bereiche im Bild leichter und besser erkennbar machen soll.

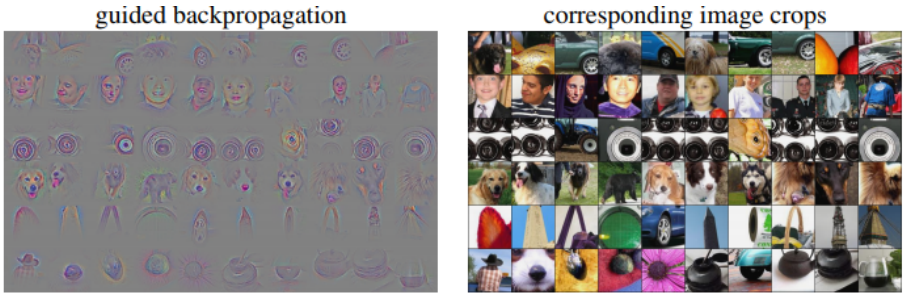


Fig. 2. Die Guided Backpropagation ist eine Erweiterung der Gradients-Methode und liefert entsprechend feinere Erklärungen.

Activation Maximization Einer der ersten Ansätze um das Innenleben von neuronalen Netzen zu visualisieren ist die *Activation maximization* [10]. Dabei wird für Klassifikationsprobleme ein grundlegender Prototyp für jede Klasse erstellt. Dafür wird eine Eingabe so optimiert, dass die Aktivierung im Netz an einer bestimmten Stelle maximiert wird. Dieser generierte Input ist dadurch eine optimale Repräsentation im Eingaberaum, mit maximalem Ausschlag für eine (oder mehrere) Klassen. Diese Technik kann auf beliebige Schichten innerhalb eines DNN angewendet werden und soll zeigen, welche Form von Eingaben vom Netz erwartet werden.

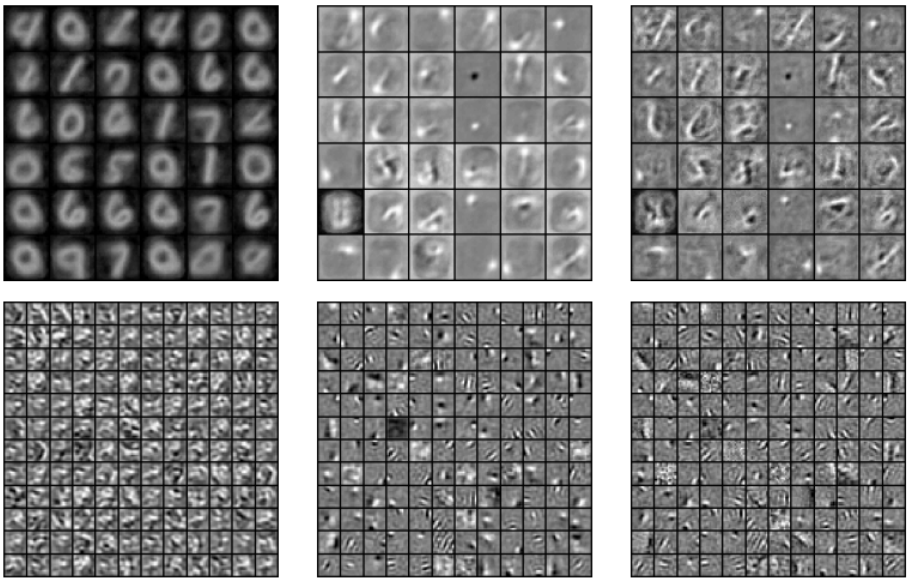


Fig. 3. Je nach Schicht ist durch die Activation Maximization erkennbar, auf welche Bereiche das DNN besonderen Wert legt.

Model-Distillation Bei den Methoden dieser Kategorie wird ein weiteres Modell trainiert. Dieses soll das Verhalten des zu untersuchenden DNN möglichst genau simulieren. Das nachgeahmte DNN wird in dem Zusammenhang als Lehrer und das simulierende Modell Schüler bezeichnet.
Bei der Model-Distillation wird also ein Schüler-Netzwerk anhand der Ausgaben des Lehrers trainiert. Dafür wird mittels bestimmter Eingabe-Daten eine Ausgabe des Lehrers erzeugt, der Schüler trainiert entsprechend mit den gleichen Daten. Das Lernziel des Schülers ist es allerdings nicht die korrekte Ausgabe zu erzeugen, sondern das Verhalten des Lehrers möglichst genau nachzuahmen, so dass die Ausgaben beider Netze im besten Fall identisch sind.
Die damit erzeugten Erklärungen sind dennoch Hypothesen, die vom Training des Schülers und der verwendeten Technik abhängig sind.
Meist werden für diesen Task Schüler-Netze verwendet, welche eine weitaus geringere Lernkapazität (weniger Parameter) als der Lehrer aufweisen, daher wird diese Überführung in ein kleineres Modell auch verwendet um Netzwerke, mit annehmbaren Einbußen in der Genauigkeits, zu komprimieren.
Es wird zwischen zwei Unterkategorien unterschieden, *lokale Approximation* und *Model Translation*. Bei der lokalen Approximation lernt ein simples Modell auf einem kleinen Subset der Daten das DNN nachzuahmen. Sind die Daten genügend nah beieinander kann ein DNN in mehrere erklärbare Netze destilliert werden, so dass z.B. für jede Klasse ein erklärbares Netz vorliegt. Während bei der Model Translation ein Schüler-Netzwerke trainiert wird, dessen Ziel es ist das

Verhalten des Lehrers global, bzw. über alle Daten des Datensatzes, nachzubilden.

Loal Interpretable Model-agnostic Explenations (LIME) Bei LIME handelt es sich um eine lokale Approximation, bei der die Merkmale der Eingabedaten identifiziert werden, die für die Zuordnung einer Instanz zu einer Klasse sprechen [20]. Dafür wird eine Eingabe iterativ verändert und die Antwort des Black-Box-Modells beobachtet. Bei Bildern werden dazu einzelne Segmente verdeckt, siehe Fig. 4. Diese veränderten Eingaben werden dann vom Modell bewertet. Iterativ wird damit ein erklärbares Modell aufgebaut, welches das Black-Box-Modell nachbildet. Mittels LIME kann dann erkannt werden, welche Bereiche

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Fig. 4. Mittels LIME können die Bildbereiche erkannt werden, welche die Klassifikation unterstützen (Verdeckung verringert Zuversicht) oder dagegen sprechen (Verdeckung erhöht Zuversicht).

der Eingabe für die Ausgabe des Modelles entscheidend waren. Die grundlegende Funktionsweise von LIME wurde von vielen modernen Techniken adoptiert und auf verschiedene Gebiete angepasst.

Tree Regularization Ein Beispiel für eine Model Translation Technik ist die Tree Regularization [27]. Hierfür wird ein Entscheidungsbaum so trainiert, dass er das Verhalten des Lehrer-Modells simulieren kann. Entscheidungsbäume sind, aufgrund ihrer Struktur, von Natur aus erklärbar. Um den Baum in eine überschaubaren Größe zu halten, wird ein Regularisierungsterm verwendet, der die mittlere Weglänge (APL) innerhalb des Baumes mit einkalkuliert. Die mittlere Weglänge des Entscheidungsbaumes abzufragen verlängert den Trainingsprozess allerdings erheblich. Die Autoren trainieren daher ein zweites Modell (Surrogate) mit der Intention die mittlere Weglänge basierend auf den Modell-Parametern vorherzusagen zu können.
In der Praxis ist es schwer einen globale Entscheidungsbaum für das Netz zu erstellen, der auch gleichzeitig für Menschen verständlich ist und weiterhin eine annehmbare Genauigkeit bei der Ausgabe hat [28]. Es existiert ein Trade-Off zwischen Baum-Komplexität und Fehlerrate, ein komplexer Baum hat eine geringere

Fehlerrate, ist allerdings auch weniger verständlich für Menschen. Das Ergebnis der Technik ist stark von den gewählten Hyperparametern abhängig, so dass unterschiedliche Startpunkte zu unterschiedlichen Bäumen führen, was die Tree-Regularization unzuverlässig macht.
[28] unterteilt die Problem-Domain daher in R Bereiche, welche nicht zwangsläufig vollständig voneinander getrennt sein müssen, den Eingaberaum aber vollständig abdecken sollten. Für diese Unterteilung sollten idealerweise Domain-Experten herangezogen werden, oder alternativ entsprechend Clustering-Algorithmen verwendet werden. Für jede der Regionen wird dann ein eigener Baum trainiert. Jeder Baum benötigt dafür wiederum ein Surrogate-Model zum Vorhersagen der APL, was die Komplexität des Trainingsprozesses um den Faktor R erhöht.
Die Unterteilung der Domain in Bereiche ist allerdings nicht immer möglich oder fair. Die Hauptanwendung dieser Technik liegt vor allem im medizinischen Sektor und bei der Klassifikationproblem.

Intrinisische Methoden Statt über ein weiteres Modul in der KI-Pipeline einer Erklärung für die Netzwerkausgabe bereitzustellen können mittels intrinsischer Methoden DNN trainiert werden, die selbst eine Erklärung liefern. Dies kann dabei ein Teilproblem der der Lernaufgabe sein, so dass die Erklärung zum Teil der Ausgabe wird oder über die Netzwerkstruktur abgelesen werden. Die Ansätze werden vor allem zwei Kategorien eingeteilt, *Attention Mechanisms* und *Joint Training*. Bei Attention Mechanisms wird die Verteilung der Eingabe gelernt und ein entsprechender Gewichtsvektor zur weiteren Verarbeitung ausgegeben. Das Joint Training generiert eine Erklärung als Teil der Trainingsaufgabe, dafür müssen allerdings Erklärungen erstellt werden, welche dann wiederum teil der Eingaben beim Trainieren sind.

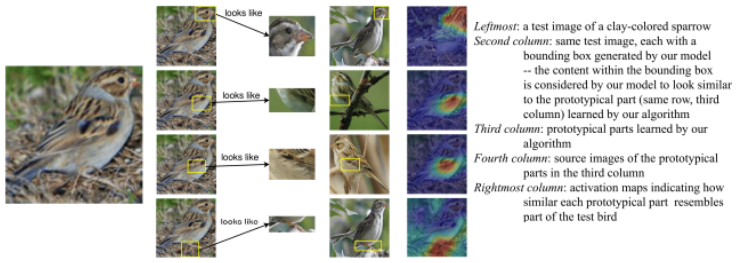


Fig. 5. Das Bild des Spatzen wird mit unterschiedlichen Prototypen aus dem Datensatz assoziiert.

Model Prototype Das ProtoPNet [7], besteht aus einem Convolutional Neural Network (CNN) zur Feature-Extraktion und einem Prototyp-Layer. Dieser berechnet die Distanz der Features zu den Prototypen aus dem Datenset. Bei

einem Prototyp handelt es sich um Repräsentationen einer speziellen Klasse. Anhand dieser Prototypen können dann die Bereiche des Bildes zugeordnet werden, die für das Netz mit bestimmten Prototypen in Verbindung gebracht werden, siehe Fig. 5.

Weitere Bereiche Neben den vorgestellten Kategorien gibt es weitere Bereich deren Ziel zum gleichen Maße das Verständnis des trainierten DNN ist um eventuelle Fehlerquellen ausfindig zu machen. Es gibt eine Reihe von Techniken welche die Evolution eines Modelles während des Trainingsprozess visualisieren. Diese werden in erster Linie dazu verwendet um zu studieren, nach welchen Prinzipien die Gewichte des DNN angepasst werden. Der Schwerpunkt liegt oft dennoch auf unterschiedlichen Aspekten:

- [19] untersuchen den Entwicklungsprozess von Schichten während des Trainings
- [19, 30] untersuchen die Konvergenz von Schichten
- [1] untersucht die Generalisierungseigenschaften von DNN

3.2 Verifyable AI

Eine menschenverständliche Erklärung einer KI-Ausgabe ist dennoch keine formelle Beschreibung und schwer zu verallgemeinern. Traditionell erfolgen KI-Tests über eine große Menge von Datenpunkten im Eingaberaum, die tatsächlichen Ausgaben des Netzwerkes werden dann mit gewünschten Ausgaben abgeglichen, so dass eine prozentuale Genauigkeit (oder andere Metrik) für das Netz mit den gegebenen Daten die Performance der KI beschreibt. In den meisten Problemfällen, wie auch bei den bahnspezifischen Beispielen, ist der Eingaberaum jedoch unendlich groß. Eine komplette Abdeckung der Tests ist daher praktisch nicht möglich. Weiterhin wurde mit der Entdeckung von Adversarial Examples (Sektion 4.1) gezeigt, dass selbst Netze mit außerordentlich geringen Fehlerraten durch manipulierte Eingaben zu falschen Ergebnissen kommen.
Es gibt öffentlich zugängliche Repositories, welche Werkzeuge zur Verfügung stellen um das Verhalten von KI zu verifizieren.

VerifAI Das VerifAI-Toolkit [9] führt eine Analyse basierend auf einem Simulator, sowie dessen Umgebungsbeschreibung durch. Das Toolkit enthält dabei eine Reihe von Techniken (Fuzz-testing, Counterexamples, Datenaugmentation, etc.), die entsprechend genutzt werden um die KI zu analysieren und zu bewerten. VerifAI wird in der Praxis verwendet um eine formale Analyse von KI-Systemen durchzuführen [11, 12].

CNN-Cert Eine weitere Bibliothek zur Verifikation Netzwerken ist CNN-Cert [4] von IBM. Dieses berechnet eine untere Grenze der minimalen Verzerrung einer Eingabe als Robustsheitsgarantie. Die Technik soll auch auf Netze mit hoher Kapazität (z.B. ResNet) anwendbar sein und unterstützt allgemein Convolutional Neural Networks (CNN).

3.3 Nachweis über die gelernten Features

Tiefe DNN verfügen teilweise über Millionen von Parametern, welche nach dem Training eine Art "Footprint" für das Netz bilden. In der Literatur konnte allerdings kein Ansatz gefunden werden, der eine solche verifikation beschreibt. Die Bereiche der *Semantic Correspondence* und *Image Captioning* haben ähnliche Grundprobleme, jedoch handelt es sich bei der Semantic Correspondence um Techniken zum Matchen von Punkten verschiedener Instanzen einer Klasse, was wiederum für *Image Matching* und *Gemoetric Aligment* verwendet wird. Es handelt sich aber um keine Beschreibung der antrainierten Features, sondern um eine Beschreibung von Datenpunkten von Instanzen. Das Image Captioning wird verwendet um Bildbeschreibungen zu generieren und ist ebenfalls für Instanzen, nicht für globale Features geeignet.
In der Literatur findet sich keine Möglichkeit eine semantische Beschreibung eines Features zu generieren.

4 Angriffspunkte

Die Robustheit von KI-Systeme gegenüber von gezielten Angriffen ist ein wichtiger Bestandteil der Sicherheitsbedenken. Die Gründe für solche Angriffe von externen Stellen sind vielzählig und werden an dieser Stelle nicht weiter erläutert. Das Ziel eines Angriffs auf ein KI gestütztes System ist das provozieren eines defekten Zustanden bzw. einer unerwarteten Ausgabe. Im weiteren Sinne kann das Ziel der Angreifer auch das Erreichen eines bestimmten Zustandes sein, indem die KI zu einer vorher festgelegten Ausgabe geführt wird.
In Systemen mit offenen Eingaben, bei denen nicht garantiert werden kann, dass keine manipulierte Daten ins Netz gelangen, kann nicht von einer sicheren Umgebung ausgegangen werden. Die Risiken von erfolgreichen Angriffen hängen vor allem vom Einfluss der KI auf das Gesamtsystem ab. Einige Extremfälle für die bahnspez. Beispiele sind in Tabelle 3 aufgeführt.
In einem KI-System gibt es drei Punkte an denen Angreifer potentiell Einfluss auf dieses gewinnen können. Durch direkte Manipulation der physischen Sensoren oder der Datenübertragung sind Angreifer in der Lage direkt in die Datensammlung bzw. -transmission einzugreifen. Dieser Aspekt ist allerdings immer gegeben wenn physische Apparaturen verwendet werden, zu denen Menschen Zugriff haben könnten und ist daher nicht KI-spezifisch.
Ein weiterer Ansatzsatzpunkt für Angreifer sind die Daten selbst. Wenn der Angreifer Wissen über die Datensammlung hat, kann er aktiv manipulierte Samples in das Training mit aufnehmen. Diese Art von Angriff wird auch als *Poisoning Attack* bezeichnet. Für das Projekt werden diese Angriffe allerdings ausgegliedert, da die Datensammlung nicht Teil der Sicherheit des KI-Systems ist.
Die dritte Art von Angriff wird auch als *Evasion Attack* bezeichnet. Diese fasst eine Kategorie von Angriffen zusammen, bei der KI-System zur Produktionszeit, bzw. nach abgeschlossenem Training mit speziell angepassten Samples übergeben werden. Das Ziel des Angriffs ist eine falsche Vorhersage des DNN

herbeizuführen. Diese manipulierten Eingaben werden auch als *Adversarial Examples* bezeichnet.

Table 3. Mögliche Szenarien im Falle eines erfolgreichen Angriffs auf das KI-System innerhalb einer Bahn.

Bahnspez. Beispiel	Auswirkungen
Beispiel1	• Schäden am Zug • Entgleisung
Beispiel 2	• Schäden am Zug • Falsche Fehlermeldungen
Beispiel 3	• Falsche Berechnung der Ankunftszeit • Zugkollision

4.1 Adversarial Examples

Unter Adversarial Examples (AE) wird eine Attacke verstanden, bei der ein Input so abgeändert wird, dass ein menschlicher Betrachter weiterhin eine korrekte Einordnung des Samples vornehmen könnte, eine KI jedoch eine falsches Ergebnis berechnet, obwohl dieses im unbearbeiteten Zustand potentiell richtig zugeordnet werden könnten. Selbst Netzwerke, welche auf einer großen Menge von Daten trainiert wurden und normalerweise eine genauso geringere Fehlerquote wie Domain-Experte vorweisen können sind von solchen Angriffen betroffen. Die Gründe warum solche Attacken möglich sind, sind nicht eindeutig auflösbar. Die Begründer der Technik beschreiben selbst, dass ein DNN, egal wie gut dessen Performance auf dem Testdatensatz ist, nicht wirklich die unterliegende Struktur der Daten lernt [13]. Ein Grund für diese Schwachstelle könnte die Linearität von Modellen sein. Oft wird die euklidische Distanz mit der wahrnehmbaren Distanz gleichgesetzt, mit der Annahme, dass Samples mit sehr geringer Distanz gleich eingeordnet werden. In Wirklichkeit können kleine vom Menschen nicht erkennbare Änderungen große Auswirkungen innerhalb des Netzwerkes haben. Über diese Anpassungen kann ein Modell dazu gebracht werden ein Signal exklusiv zu betrachten, da dieses am besten nach den Gewichten ausgerichtet ist, obwohl es vielleicht größere Ausschläge gebe [13].
[13] zeigen dass AE anhand eines simplen Algorithmus generiert werden können. Darauf basierend wurde eine Reihe von Techniken entwickelt um immer stärkere AE zu generieren, was auch weiterhin der Gegenstand aktueller Forschung ist. Man unterscheidet bei den aufgeführten Methoden zwischen zielgerichtet und nicht-zielgerichtet. Die nicht-zielgerichteten Methoden haben die Aufgabe eine beliebige falsche Ausgabe zu erzeugen, während zielgerichtete Angriffe eine bestimmte Klasse bzw. Ausgabe herbeiführen sollen.
Weiterhin wird unterschieden zwischen Einzelschritt- und iterativen Methoden. AE, welche mit Einzelschrittmethoden generiert wurden, führen oft auch bei anderen Modellen zu einer falschen Klassifizierung. Während AE aus iterativen

Methoden zwar weitaus effektiver auf dem Ausgangsmodell sind, jedoch meist nicht auf andere Modelle transferierbar sind, was sie für Black-Box-Angriffe [8] ungeeignet mach. Von Transferierbarkeit spricht man, wenn ein AE, welches mittels eines bestimmten DNN generiert wurde auch auf anderen DNN mit dem gleichen Effekt angewendet werden kann.

Fast gradient sign method (FGSM) Die FGSM gehört zu den ersten white-box Methoden um AEs zu generieren und kann einfach verwendet werden um die Robustheit eines DNN zu überprüfen. Dafür wird ein Vektor η , basierend auf einer Eingabe x und einer Zielausgabe y_{true}, sowie den Modellparametern θ berechnet. der Faktor ϵ ist die kleinste vom Sensor wahrnehmbare Änderung, bei Bildern könnte dies z.B. $1/255$ sein. Dieser Wert sollte möglichst gering gehalten werden, damit die eingeführten Störungen nicht mit bloßem Auge sichtbar sind.

$$\eta = \epsilon \cdot sign(\nabla_x J(\theta, x, y_{true})) \tag{1}$$

Mit der Kostenfunktion J kann dann die Störung η berechnet werden. Ein AE kann dann einfach mittels $\eta + x$ generiert werden. Die *sign*-Methode setzt lediglich alle Werte unter 0 auf -1 und alles drüber auf 1, der Wert 0 selbst bleibt unverändert.
AE sprechen gezielt Schwachstellen in den Gewichten von DNN an. Es ist daher wie vorher angenommen, nicht notwendig die gestörten Bilder direkt ans das Netzwerk zu übergeben. Bilder in die mit den entsprechenden Techniken Störungen eingebaut wurden, werden auch nach dem Ausdrucken und erneut fotografieren weiterhin falsch klassifiziert [16]. Dadurch ist es, sind zumindest theoretisch, sehr einfach Angriffe in der echten Welt durchzuführen. [6] stellen einen Angriff vor, bei dem ein bestimmtes Muster ausgedruckt wird, welches dann in der nähe von zu klassifizierenden Objekten abgelegt wird. Die getesteten DNN waren dadurch nicht mehr in der Lage diese korrekt zu klassifizieren.

Basic iterative method (BIM / I-FGSM) Die Effektivität des FGSM kann durch Iteration weiter gesteigert werden. Der Algorithmus wird um den Faktor α erweitert, welcher die maximale Änderung für einzelne Pixel innerhalb einer Iteration festschreibt, die Autoren [16] setzen hier einen Wert von 1 an. Die ermittelte Störung wird dann in jeder Iteration zum Original addiert und entsprechend des Wertebereichs abgeschnitten.

$$x_{N+1}^{adv} = clip\{x_N^{adv} + \alpha sign(\nabla_x J(\theta, x_N^{adv}, y_{true}))\} \tag{2}$$

Bei der hier aufgeführten nicht zielgerichteten Methode wird ein Bild iterativ so angepasst, dass sich der Fehler des Netzwerkes gegenüber dem eigentlichen Label/Ausgabe erhöht. Die Signifikanz von falschen Vorhersagen kann daher aber gering ausfallen, wenn beispielsweise statt einer bestimmten Hunderasse eine andere Hunderasse erkannt wird.
Bei der zielgerichteten Variante wird der Fehler des Netzes bezüglich der Zielausgabe entsprechend minimiert. Um einen maximalen Fehler herbeizuführen kann daher die *Iterative least-likely class mehtod* (Iter-LL) angewendet werden. Diese

sucht zuerst die Klasse aus der Netzwerkausgabe, welche für das Netz um unwahrscheinlichsten ist und versucht dann mittels I-FGSM denn Input so zu stören, dass diese Klasse mit möglichst hoher Zuversicht vorausgesagt wird.

$$x_{N+1}^{adv} = clip\{x_N^{adv} - \alpha sign(\nabla_x J(\theta, x_N^{adv}, y_{LL}))\} \tag{3}$$

Mittels Iter-LL können Eingaben so sehr gestört werden, dass die Genauigkeit des Netzes gegen 0% geht. Die durchgeführten Änderungen in der Eingabe sind für den Menschen nicht wahrnehmbar. Bei Durchführen des FGSM mit hohem ϵ geht die Genauigkeit der Modellvorhersage ebenfalls unweigerlich gegen 0%, allerdings sind die AE dann auch für den Menschen nicht mehr erkennbar.
[16] konnten ebenfalls zeigen, dass nur AE, welche mit FGSM generiert wurden auch dann zuverlässig falsch Aussagen provozieren, wenn Bildoperationen angewendet werden. Der I-FGSM, sowie Iter-LL verlieren ihre Adversary Eigenschaften teilweise durch weitere Bildoperationen. Die Autoren vermuten, dass die iterativen Methoden sehr feine Änderungen durchführen, welche stark auf das Bild angepasst sind, während der FGSM gröbere Änderungen vornimmt. Es existieren Erweiterungen dieser Methode, wie die Momentum iterative FGSM (MI-FGSM), welche aktuell eine der effektivsten Methoden zur Generierung von AE ist [8].

4.2 Robustes Training

Es existieren viele Versuche die Angriffe durch AE abzuwehren. Allgemein gilt, dass Netzwerke mit einer hohen Kapazität bereits universell robuster sind als kleinere Modelle [17]. Um DNN gegenüber Adversarial Attacks (AA) robust zu machen wird beim Trainingsvorgang jedes Sample zu einem AE transformiert. Dieses sogenannte Adversarial Training (AT) bringt eigene Probleme mit sich. Wenn das trainierende Modell verwendet wird um die AE zu generieren, lernt dieses nicht zwangsläufig sich besser zu verteidigen, sondern lernt selbst schlechter Angriffe durchzuführen. Das heißt AE die mit diesem Netz generiert wurden können mit hoher Wahrscheinlichkeit von anderen Netzen richtig klassifiziert werden [24]. Dieser Effekt wird auch als *Reward Hacking* bezeichnet. Der Agent lernt ein Verhalten, welches den Gewinn maximiert, was aber nicht die Intention des Architekten widerspiegelt.
Dieses Problem kann umgangen werden, indem das Generieren der AE beim Trainingsprozess nicht vom trainierenden Netzwerk übernommen wird, sondern von einem unabhängigen, gesondert vortrainierten DNN. Dadurch können Black-Box-Angriffe realitätsnah abgebildet werden.

Die Existenz von AEs ist nicht nur ein Beiprodukt der DNN, sondern kann ebenfalls auf die Daten zurückgeführt werden [14]. Netzwerke lernen während des Trainings robuste, sowie nicht robuste Features. Die Störungen eines AE greifen gezielt die nicht-robusten Features an um ein falsches Ergebnis zu provozieren. AE können auch zu anderen Modellen transferiert werden, da selbst unterschiedliche DNN die gleichen nicht-robusten Features lernen [14].
Bei einer Vorhersage existieren eine Reihe von nützlichen Features, die mit einer

bestimmten Distribution mit dem korrekten Label korrelieren. Robuste nützliche Features, bleiben auch durch adversere Störungen nützlich. Die Korrelation von nützlichen nicht-robuste Features kann durch AA umgekehrt werden, so dass eine Antikorrelation vorgetäuscht wird [14]. Beim normalen Trainingsvorgang versucht das Modell mit allen Mitteln einen besseren Fehlerwert zu erreichen und nutzt daher jegliche nützlichen Features.
Beim angepassten robusten Training wird eine Unterscheidung zwischen robusten und nicht-robusten Features getroffen, mit dem Ziel, dass ein DNN nur nützliche robuste Features lernt. Das Training auf nicht-robusten Features führt zu einer hohen Genauigkeit, aber auch zu einer geringen Robustheit gegenüber AEs, was zeigt, dass moderne DNN genau solche Features benötigen um möglichst akkurat zu sein.

Um sehr performante Modelle zu trainieren ist es notwendig nicht-robuste Features zu lernen, was jedoch in einem angreifbaren DNN resultiert [25]. Robuste Modelle lernen eine Repräsentation der Daten, welche gegenüber visuell schwache Störungen invariant ist, dementsprechend sind robuste Modell eher im Einklang mit der menschlichen Wahrnehmung. Vergleicht man die visualisierten Gradienten von robusten Modellen gegenüber nicht-robusten Modellen, fällt schnell auf, dass die erkannten Features den Bildpunkten entsprechen, die auch für Menschen notwendig sind um ein Objekt entsprechend zu klassifizieren, während die Gradienten nicht-robuster Modelle eher an Noise, ohne auffallige Muster welche auf das Ausgangsobjekt schließen lassen würdne, erinnern (siehe Fig. 6).

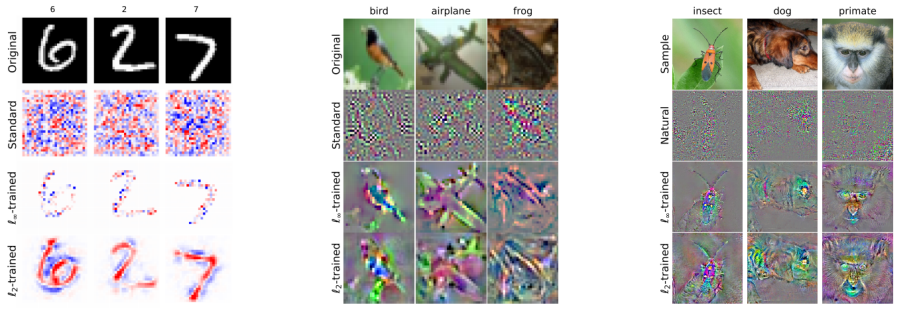


Fig. 6. Die von robusten Netzwerken erkannten Features, spiegel die menschliche Wahrnehmung wider, während die Gradienten von nicht-robusten Netzwerken für Menschen unkenntlich sind.

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memo-

rization in deep networks. In: International Conference on Machine Learning. pp. 233–242. PMLR (2017)
2. Birdsall, M.: Google and ite: The road ahead for self-driving cars. Institute of Transportation Engineers. ITE Journal **84**(5), 36 (2014)
3. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K.: End to end learning for self-driving cars (2016). https://doi.org/10.48550/ARXIV.1604.07316, https://arxiv.org/abs/1604.07316
4. Boopathy, A., Weng, T.W., Chen, P.Y., Liu, S., Daniel, L.: Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3240–3247 (2019)
5. Broggi, A., Buzzoni, M., Debattisti, S., Grisleri, P., Laghi, M.C., Medici, P., Versari, P.: Extensive tests of autonomous driving technologies. IEEE Transactions on Intelligent Transportation Systems **14**(3), 1403–1415 (2013). https://doi.org/10.1109/TITS.2013.2262331
6. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
7. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
8. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
9. Dreossi, T., Fremont, D.J., Ghosh, S., Kim, E., Ravanbakhsh, H., Vazquez-Chanlatte, M., Seshia, S.A.: Verifai: A toolkit for the formal design and analysis of artificial intelligence-based systems. In: International Conference on Computer Aided Verification. pp. 432–442. Springer (2019)
10. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. University of Montreal **1341**(3), 1 (2009)
11. Fremont, D.J., Chiu, J., Margineantu, D.D., Osipychev, D., Seshia, S.A.: Formal analysis and redesign of a neural network-based aircraft taxiing system with verifai. In: International Conference on Computer Aided Verification. pp. 122–134. Springer (2020)
12. Fremont, D.J., Kim, E., Pant, Y.V., Seshia, S.A., Acharya, A., Bruso, X., Wells, P., Lemke, S., Lu, Q., Mehta, S.: Formal scenario-based testing of autonomous vehicles: From simulation to the real world. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–8. IEEE (2020)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. arXiv preprint arXiv:1905.02175 (2019)
15. Kato, S., Takeuchi, E., Ishiguro, Y., Ninomiya, Y., Takeda, K., Hamada, T.: An open approach to autonomous vehicles. IEEE Micro **35**(6), 60–68 (2015). https://doi.org/10.1109/MM.2015.133
16. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world (2016)
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
18. Muller, U., Ben, J., Cosatto, E., Flepp, B., Cun, Y.: Off-road obstacle avoidance through end-to-end learning. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.)

Advances in Neural Information Processing Systems. vol. 18. MIT Press (2005), https://proceedings.neurips.cc/paper/2005/file/fdf1bc5669e8ff5ba45d02fded729feb-Paper.pdf
19. Raghu, M., Gilmer, J., Yosinski, J., Sohl-Dickstein, J.: Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. arXiv preprint arXiv:1706.05806 (2017)
20. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
21. Sallab, A.E., Abdou, M., Perot, E., Yogamani, S.: Deep reinforcement learning framework for autonomous driving. Electronic Imaging **29**(19), 70–76 (jan 2017). https://doi.org/10.2352/issn.2470-1173.2017.19.avm-023, https://doi.org/10.2352%2Fissn.2470-1173.2017.19.avm-023
22. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
23. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
24. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017)
25. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018)
26. Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, R., Clark, M., Dolan, J., Duggins, D., Galatali, T., Geyer, C., et al.: Autonomous driving in urban environments: Boss and the urban challenge. Journal of Field Robotics **25**(8), 425–466 (2008)
27. Wu, M., Hughes, M.C., Parbhoo, S., Zazzi, M., Roth, V., Doshi-Velez, F.: Beyond sparsity: Tree regularization of deep models for interpretability. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
28. Wu, M., Parbhoo, S., Hughes, M., Kindle, R., Celi, L., Zazzi, M., Roth, V., Doshi-Velez, F.: Regional tree regularization for interpretability in deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6413–6421 (2020)
29. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. CoRR **abs/1906.05113** (2019), http://arxiv.org/abs/1906.05113
30. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)

Lehrstuhl Privatrecht und Recht des geistigen Eigentums
Prof. Dr. Dagmar Gesmann-Nuissl

Autoren: Dagmar Gesmann-Nuissl / Stefan Kunitz

AP 2 – Rechtlicher Rahmen: Zulassungsprozess

Der abstrakte Zulassungsprozess

Das Inverkehrbringen von Fahrzeugen sowie die Inbetriebnahme von Eisenbahninfrastruktur bedürfen regelmäßig einer **Inbetriebnahmegenehmigung**.

(a) Genehmigungserfordernis

Die Erteilung von Inbetriebnahmegenehmigungen ist in der Verordnung über die Erteilung von Inbetriebnahmegenehmigungen für das Eisenbahnsystem (Eisenbahn-Inbetriebnahmegenehmigungsverordnung – EIGV)[1] geregelt; das Erfordernis einer Inbetriebnahmegenehmigung ist in § 9 EIGV festgelegt.

Die durch das Bundesministerium für Verkehr und digitale Infrastruktur erlassene Rechtsverordnung (Ermächtigung in § 26 Abs. 1 Nr. 1 AEG), dient der Umsetzung des vierten Eisenbahnpaketes der EU in nationales Recht, insbesondere in Umsetzung der Richtlinie (EU) 2016/797 (sog. Interoperalitätsrichtlinie; kurz: RL (EU) 2016/767).

Für die Zulassung bzw. Inbetriebnahmegenehmigung nach dem EIGV muss deren Anwendungsbereich eröffnet sein. Er ist in § 1 Abs. 3 EIGV geregelt und erstreckt sich auf das regelspurige Eisenbahnsystem im Zuständigkeitsbereich des Eisenbahn-Bundesamtes, d.h. auf alle Eisenbahnen des Bundes (§ 4 Abs. 6 AEG, § 3 BEVVG).

Sofern eine Zulassung bzw. Inbetriebnahmegenehmigung im Anwendungsbereich nach § 9 EIGV begehrt wird, muss das entsprechende System einen **vorgegebenen Zulassungsprozess** durchlaufen. Wie dieser Prozess im Einzelnen ausgestaltet ist, richtet sich danach, für welches Teilsystem eine Zulassung angestrebt wird.

(b) Teilsysteme

Die Aufteilung in die verschiedenen Systeme erfolgt nicht in der EIGV selbst. Sie beschreibt nur die Genehmigungsverfahren für die verschiedenen Teilsysteme.

[1] Richtlinie (EU) 2016/797 des Europäischen Parlaments und des Rates vom 11. Mai 2016 über die Interoperabilität des Eisenbahnsystems in der Europäischen Union, Amtsblatt der Europäischen Union L 138 vom 26. Mai 2016 sowie deren Berichtigungen v. 17.9.2020, ABl. L 303/23 sowie v. 22.12.2021, ABl. L 458/539.

Vielmehr wird in § 2 Nr. 26 EIGV für die Definition eines Teilsystems auf den Anhang II der RL (EU) 2016/797 verwiesen. Dort wird das Gesamtsystem „Eisenbahnsystem der Union“ in **strukturelle und funktionelle Teilsysteme** untergliedert.

Strukturell in die Bereiche:

- **Infrastruktur**: Unter Infrastruktur fallen dabei Gleise, Weichen, Bahnübergänge, Kunstbauten (Brücken, Tunnel usw.), eisenbahnbezogene Bahnhofsbestandteile (u. a. Eingänge, Bahnsteige, Zugangs- und Servicebereiche, Toiletten und Informationssysteme sowie deren Zugänglichkeitsfunktionen für behinderte Menschen und Personen mit eingeschränkter Mobilität), Sicherheits- und Schutzausrüstung.
- **Energie**: Das Teilsystem Energie umfasst Energieversorgungssystem, einschließlich Oberleitungen und streckenseitiger Teile der Einrichtungen zur Messung und Abrechnung des Stromverbrauchs.
- **Strecken- und fahrzeugseitige Zugsteuerung/Zugsicherung und Signalgebung**: Vom Bereich der fahrzeugseitigen/streckenseitigen Zugsteuerung, Zugsicherung und Signalgebung (ZZS) werden alle erforderlichen fahrzeugseitigen/streckenseitigen Ausrüstungen zur Gewährleistung der Sicherung, Steuerung und Kontrolle der Bewegung von Zügen, die zum Verkehr im Netz zugelassen sind, erfasst.
- **Fahrzeuge**: Unter das Teilsystem Fahrzeuge fallen Wagenkastenstruktur, System der Zugsteuerung und Zugsicherung sowie die dazugehörigen Einrichtungen des Zuges, Stromabnahmeeinrichtungen, Traktions- und Energieumwandlungseinrichtungen, fahrzeugseitige Einrichtungen zur Messung und Abrechnung des Stromverbrauchs, Bremsanlagen, Kupplungen, Laufwerk (Drehgestelle, Achsen etc.) und Aufhängung, Türen, Mensch-Maschine-Schnittstellen (Triebfahrzeugführer, Zugbegleitpersonal und Fahrgäste, einschließlich Zugänglichkeitsfunktionen für behinderte Menschen und Personen mit eingeschränkter Mobilität), passive oder aktive Sicherheitseinrichtungen und Erfordernisse für die Gesundheit der Fahrgäste und des Zugbegleitpersonals.

Funktionell in die Bereiche:

- **Betriebsführung und Verkehrssteuerung**: Betriebsführung und Verkehrssteuerung sind die Verfahren und zugehörige Ausrüstungen, die eine kohärente Nutzung der verschiedenen strukturellen Teilsysteme erlauben, und zwar sowohl im Normalbetrieb als auch bei Betriebsstörungen, einschließlich insbesondere der Zugbildung und Zugfahrten, der Planung und der Abwicklung der Betriebsführung. Die Gesamtheit der erforderlichen beruflichen Qualifikationen für die Durchführung von Schienenverkehrsdiensten jeglicher Art.
- **Instandhaltung**: Das Teilsystem Instandhaltung umfasst Verfahren, zugehörige Ausrüstungen, logistische Instandhaltungseinrichtungen, Reserven zur Durchführung vorgeschriebener Instandsetzungsarbeiten und vorbeugender Instandhaltung im Hinblick auf die Gewährleistung der Interoperabilität des Eisenbahnsystems der Union und der erforderlichen Leistungsfähigkeit.
- **Telematikanwendungen für den Personen- und Güterverkehr**: Das Teilsystem der Telematikanwendungen umfasst gemäß Anhang I zur Richtlinie (EU) 2016/797 zwei Bereiche: Anwendungen im Personenverkehr, einschließlich der Systeme zur Information der Fahrgäste vor und während der Fahrt, Buchungssysteme, Zahlungssysteme, Reisegepäckabfertigung, Anschlüsse zwischen Zügen und zwi-

schen der Eisenbahn und anderen Verkehrsträgern. Zudem Anwendungen im Güterverkehr, einschließlich der Informationssysteme (Verfolgung der Güter und der Züge in Echtzeit), Rangier- und Zugbildungssysteme, Buchungssysteme, Zahlungs- und Fakturierungssysteme, Anschlüsse zu anderen Verkehrsträgern, Erstellung elektronischer Begleitdokumente.

Für die Erlangung einer Zulassung nach EIGV ist vor allem maßgeblich, ob das zuzulassende System dem Teilsystem Fahrzeuge im nachfolgendem dem „mobilen Teilsystem“ (i) oder einem der anderen Teilsysteme des Anhang II der RL (EU) 2016/797, im Folgendem „sonstiges Teilsystem“ (ii) zuzurechnen ist.

Im ersteren Fall ist ein Zulassungsverfahren für das erstmalige Inverkehrbringen nach § 9 Abs. 1 EIGV, für das Inverkehrbringen eines aufgerüsteten Fahrzeugs nach § 9 Abs. 3 EIGV erforderlich. Für die sonstigen Teilsysteme sind die Absätze § 9 Abs. 2 und 4 EIGV einschlägig, auch hier je nachdem ob ein erstmaliges Inverkehrbringen oder eine Aufrüstung erfolgen soll. Die Verortung der Genehmigungserfordernisse für Bestandteile der nicht mobilen Teilsysteme im Teil der EIGV, der sich maßgeblich den Zulassungserfordernissen für das mobile Teilsystem widmet, ist dabei aus gesetzessystematischer Sicht leider wenig glücklich.

Dabei ist ein erstmaliges Inverkehrbringen, unabhängig welches Teilsystem in Rede steht, immer genehmigungsbedürftig; ein Inverkehrbringen nach einer Aufrüstung lediglich, wenn eine in Anlage 4 zur EIGV genannte Maßnahme durchgeführt werden soll. Maßnahmen die unter Anlage 5 zur EIGV fallen, sind als reine Austauschmaßnahmen von dem Genehmigungserfordernis explizit ausgenommen.

(i) Mobiles System (Zulassungsprozess, Erlangen d. Fahrzeugzulassung)

Dem „mobilen System“ sind gem. Anhang II der RL (EU) 2016/797 Wagenkastenstruktur, zugseitige ZZS, Stromabnahmeeinrichtungen, Traktions- und Energieumwandlungseinrichtungen, fahrzeugseitige Stromverbrauchsmess- und Ladeeinrichtungen, Bremsanlagen, Kupplungen, Laufwerk (Drehgestelle, Achsen etc.) und Aufhängung, Türen, Mensch-Maschine-Schnittstellen (Triebfahrzeugführer, Zugbegleitpersonal und Fahrgäste, einschließlich Zugänglichkeitsfunktionen für behinderte Menschen und Personen mit eingeschränkter Mobilität), passive oder aktive Sicherheitseinrichtungen und Erfordernisse für die Gesundheit der Fahrgäste und des Zugbegleitpersonals zuzurechnen. Zusammengefasst also alle Systeme die im Zug „mobil“ sind. Diese werden im Regelfall nicht als einzelne Systeme, sondern vielmehr als eine Komponente – das Fahrzeug – zugelassen.

Nach § 9 Abs. 1 und 3 EIGV bedürfen sowohl das erstmalige Inverkehrbringen eines Fahrzeuges (§ 9 Abs. 1 EIGV) als auch das Inverkehrbringen eines aufgerüsteten oder erneuerten Fahrzeuges, soweit eine Maßnahme nach Anlage 4 zur EIGV vorgenommen wurde (§ 9 Abs. 3 EIGV), einer Genehmigung (vgl. oben (a)).

1) Zuständigkeit

Zuständige Stelle für die Erteilung dieser Genehmigungen ist nach § 10 Abs. 2 EIGV die Eisenbahnagentur der Europäischen Union (ERA), soweit sich das Verwendungsgebiet des Fahrzeuges in mehreren Mitgliedsstaaten befindet. Wenn sich das Verwendungsgebiet lediglich auf einen Mitgliedsstaat erstrecken sollte, kann der Antragsteller auswählen, ob die ERA oder die jeweilige nationale Sicherheitsbehörde als Genehmigungsstelle fungieren sollen (Art.21 Abs. 8 RL (EU) 2016/797). Soweit sich das Verwendungsgebiet auf die Bundesrepublik Deutschland beschränken sollte, kann also auch das EBA als nationale Sicherheitsbehörde (§ 5 AEG, § 3 BEVVG) an die Stelle der Genehmigungsbehörde treten § 10 Abs. 2 Satz 2 EIGV.

2) Verfahren

Das Verfahren zum Erhalt einer Inbetriebnahmegenehmigung ist in der Durchführungsverordnung (EU) 2018/545 (kurz: DVO (EU) 2018/545; GIF[2]) geregelt, auf die auch in § 11 EIGV verwiesen wird. Missverständlicher Weise verweist dieser nicht auf die Genehmigung der Erneuerung und Aufrüstung, die in der Durchführungsverordnung ebenfalls geregelt sind, sondern bezieht sich lediglich auf die Inbetriebnahme von Fahrzeugen und die Fahrzeugtypgenehmigung, differenziert also nicht – wie noch § 9 EIGV – zwischen der erstmaligen Inbetriebnahmegenehmigung und der zur Aufrüstung oder Erneuerung.

Unter der Genehmigung für das Inverkehrbringen versteht man nach Art. 2 Nr. 15 der DVO (EU) 2018/545 „die Entscheidung der Genehmigungsstelle […] dass das Fahrzeug gemäß den Nutzungsbedingungen und etwaigen anderen Beschränkungen, die in der Fahrzeuggenehmigung bzw. der Fahrzeugtypgenehmigung angegeben sind, im Verwendungsgebiet in Verkehr gebracht und sicher verwendet werden kann“.

Das **Genehmigungsverfahren** an sich wird in der DVO (EU) 2018/545 in **verschiedene Phasen aufgeteilt**. Zudem werden in dieser Verordnung die Pflichten der am Verfahren Beteiligten normiert.

a) Phase 1: Vorbereitung des Antrags (Kapitel 2 der DVO (EU) 2018/545)

Kapitel 2 der DVO (EU) 2018/545 regelt die erste Phase, die **Vorbereitung des Antrages**. Dabei handelt es sich um eine rein antragstellerseitige Phase. In deren Rahmen soll der zukünftige Antragsteller die für den Antrag notwendigen Anforderungen ermitteln und die diesen entsprechenden notwendigen Dokumente beschaffen.

i) Erfassung aller anzuwendenden Anforderungen

Zunächst sind nach Art. 13 DVO (EU) 2018/545 die Anforderungen an den Antrag zu identifizieren. Hierdurch soll sichergestellt werden, dass alle während des Lebenszyklus relevanten Anforderungen in Bezug auf die Konzeption des Fahrzeuges

- ordnungsgemäß ermittelt,
- den Funktionen bzw. Teilsystemen die jeweiligen Nutzungsbedingungen oder Beschränkungen zugeordnet und
- umgesetzt und validiert werden.

[2] Genehmigung Inbetriebnahme Fahrzeuge.

Dabei sind nach Art. 13 Abs. 2 DVO (EU) 2018/545 insbesondere zu erfassen,

- die grundlegenden Anforderungen aus Art. 3 der RL (EU) 2016/797, Anhang III,
- technische Komptabilität der Teilsysteme eines Fahrzeugs,
- die sichere Integration der Teilsysteme eines Fahrzeugs und
- die technische Komptabilität des Fahrzeugs mit dem Netz im Verwendungsgebiet.

Soweit Aspekte nicht unter die technischen Spezifikationen für Interoperabilität (TSI) oder die nationalen Vorschriften (NTR) fallen, soll der Antragsteller das in der Verordnung (EU) 2013/402 (kurz: VO (EU) 2013/402) beschriebene Risikomanagementverfahren nutzen, um die grundlegenden Anforderungen in Bezug auf die Sicherheit des Fahrzeugs und der Teilsysteme sowie die sichere Integration der Teilsysteme nach Art. 13 Abs. 2 a) DVO (EU) 2018/545 zu erfassen.

ii) Bestimmung der Art der erforderlichen Genehmigung

Neben dem Erfassen der Anforderungen hat der Antragsteller zu bestimmen, **welche Art der Genehmigung** für den geplanten Einsatz des Fahrzeuges erforderlich ist. Die verschiedenen Arten der Genehmigung sind in Art. 14 DVO (EU) 2018/545 benannt:

- Die **Erstgenehmigung** als die Genehmigung einer neuen Konstruktion (Art. 14 Abs. 1 lit. a));
- die **Erneute Genehmigung**, sofern die TSI oder nationale Vorschriften modifiziert wurden und die künftigen fahrzeugtypkonformen Fahrzeuge diesen geänderten Vorschriften entsprechen müssen (Art. 14 lit. b));[3]
- die **Erweiterung des Verwendungsgebietes** betrifft bereits einen genehmigten Fahrzeugtypen, der ohne bauliche Änderung das Verwendungsgebiet erweitert (Art. 14 Abs. 1 lit c));
- die **Neue Genehmigung** für Fahrzeuge bzw. Fahrzeugtypen die auf bereits genehmigten Fahrzeugen bzw. Fahrzeugtypen basieren aber aufgrund geänderter Rahmenbedingungen eine erneute Genehmigung erfordern (Art. 14 Abs. 1 lit. d)) sowie
- die **Genehmigung auf der Grundlage eines Fahrzeugtyps** für das Inverkehrbringen von Fahrzeugen oder einer Serie solcher Fahrzeuge, die auf einem bereits genehmigten Fahrzeugtyp beruhen (Art. 14 Abs. 1 lit. e)).

Soweit ein Fahrzeugtyp genehmigt werden soll, muss der Antragsteller zuvor ermitteln, ob damit ein neuer, oder eine Variante zu einem bereits bestehenden Fahrzeugtypen genehmigt werden soll (Art. 14 Abs. 2 DVO (EU) 2018/545).

Bei Antragsstellung können die Anträge auf Erteilung einer „neuen Genehmigung“ mit einem Antrag auf „Erweiterung des Verwendungsgebietes“ oder der Antrag auf „Erstgenehmigung“ mit demjenigen zur „Genehmigung auf Grundlage eines Fahrzeugtyps“ kombiniert werden Art. 14 Abs. 3 DVO (EU) 2018/545.

[3] Leitlinien über die praktischen Modalitäten für die Fahrzeuggenehmigung, Aufl. 2018, S. 49.

iii) Änderung des bereits genehmigten Fahrzeugtyps

In Art. 15 DVO (EU) 2018/545 sind darüber hinaus die **besonderen Anforderungen** an die Genehmigung **für Änderungen an einem bereits genehmigten Fahrzeugtyp** geregelt. Auch hier muss zunächst die erforderliche Genehmigung ermittelt werden, indem geprüft wird, welche Art von Änderung vorliegt. Anschließend kann sie dann den Anforderungskategorien des Art. 15 Abs. 1 lit. a) – d) DVO (EU) 2018/545 zugeordnet werden, namentlich

- Änderungen bei denen nicht gegenüber den technischen Unterlagen abgewichen werden (Abs.1 lit. a))
- Änderungen bei denen zwar von den Unterlagen abgewichen wird, aber eine Auswirkung auf die grundlegenden Konstruktionsmerkmale nicht vorliegt und daher keine Genehmigung nach den Kriterien des Art. 21 Abs. 12 der RL (EU) 2016/797 erforderlich ist (Abs.1 lit. b)).
- Änderungen der grundlegenden Konstruktionsmerkmale für die keine Genehmigung erforderlich ist (Abs.1 lit. c)) oder
- Änderungen für die eine neue Genehmigung (Abs.1 lit. d)) nach den Kriterien des Art. 21 Abs. 12 der RL (EU) 2016/797 erforderlich ist.

Soweit nicht von den technischen Unterlagen abgewichen wird (Abs. 1 lit. a)), ist keine Überprüfung durch eine Konformitätsbewertungsstelle mehr notwendig; die ursprünglichen EG-Prüferklärungen der Teilsysteme und die Fahrzeugtypengenehmigung bleiben gültig.

Sofern eine Genehmigung nach den Kriterien des Art. 21 Abs. 12 der RL (EU) 2016/797 nicht erforderlich wird, also in den Fällen der Abs.1 lit. b) und c), sind zumindest die technischen Unterlagen zu den EG-Prüferklärungen zu aktualisieren und durch den Inhaber der Fahrzeugtypengenehmigung nach Aufforderung an die Genehmigungsstelle und/oder die zuständige nationale Sicherheitsbehörde (NSB) zu übersenden (Art. 15 Abs. 2. DVO (EU) 2018/545).

Soweit ein Fall des Abs.1 lit. c) vorliegt, also die grundlegenden Konstruktionsmerkmale geändert werden sollen, aber im Rahmen der Kriterien des Art. 21 Abs. 12 der RL (EU) 2016/797 keine neue Genehmigung erforderlich ist, hat der Inhaber der Fahrzeugtypengenehmigung nach Art. 15 Abs. 3 DVO (EU) 2018/545 eine neue Version des Typs bzw. der Typvariante zu erstellen und an die Genehmigungsstelle zu übermitteln. Diese sorgt im Anschluss für die Eintragung im ERATV[4].

Der letzte Absatz des Artikels enthält die Bestimmungen wie vorzugehen ist, wenn die Änderungsverwaltungsstelle nicht Inhaberin der Fahrzeugtypengenehmigung ist und die Änderung in eine Kategorie nach Abs. 1 lit. b) – d) fällt. In diesen Fällen ist ein neuer Fahrzeugtyp zu erstellen, der Antrag für den neuen Fahrzeugtypen durch die Änderungsverwaltungsstelle als Antragstellerin auf den bestehenden Fahrzeugtyp zu stützen und das Antragsverfahren gem. Art. 14 Abs. 1 lit. d) der DVO (EU) 2018/545 zu wählen.

[4] Das ERATV ist ein webbasiertes Register, welches von der ERA betrieben wird und über den Link https://eratv.era.europa.eu/ERATV/ erreichbar ist.

iv) Änderung des bereits genehmigten Fahrzeugs

Soweit **Änderungen an bereits genehmigten Fahrzeugen** vorgenommen werden, wird das Erfordernis einer Genehmigung ebenfalls durch Art. 15 DVO (EU) 2018/545 bestimmt.

Dabei ist zunächst zu klären, ob lediglich ein Austausch oder die Wiederherstellung mit Teilen gleicher Funktion und Leistung erfolgt. In dem klassischen Fall von Reparatur- bzw. Instandhaltungsarbeiten ist eine Genehmigung für das Inverkehrbringen nicht erforderlich.

Soweit eine anderweitige Änderung vorgenommen wird, ist diese nach Art. 15 zu prüfen und in eine entsprechende Kategorie des Art. 15 einzuordnen. Dabei ist jede Änderung im Rahmen des Konfigurationsmanagements unter der Verantwortung des Halters bzw. der von ihm beauftragten Stelle zu erfassen.

Soweit eine Änderungsverwaltungsstelle nicht Inhaberin der Fahrzeugtypgenehmigung ist und Änderungen nach Art. 15 Abs. 1 lit. b) oder c) an einem bereits genehmigten Fahrzeug zu verwalten hat, muss sie nach Art. 16 Abs. 4 DVO (EU) 2018/545

- die Abweichungen gegenüber den technischen Unterlagen, die den EG-Prüferklärungen der Teilsysteme beigefügt sind, bewerten,
- nachweisen, dass keines der in Artikel 21 Abs. 12 der RL (EU) 2016/797 genannten Kriterien erfüllt ist,
- die technischen Unterlagen, die den EG-Prüferklärungen der Teilsysteme beigefügt sind, aktualisieren und
- die Genehmigungsstelle über die Änderungen unterrichten.

Das Vorgehen kann dabei für ein Fahrzeug oder für mehrere identische Fahrzeuge erforderlich werden. Wenn eine Änderung falsch zugeordnet und/oder auf unzureichende Nachweise gestützt wurde ist die Genehmigungsstelle im Rahmen einer begründeten Entscheidung befugt, die Antragstellung binnen vier Monaten zu verlangen.

v) Ermittlung der Vorschriften einschließlich der Nichtanwendung der TSI

Bei der Vorbereitung des Antrages ist auch zu ermitteln, **ob und welche TSI einschlägig ist und welche nationalen Vorschriften Anwendung finden**. Dabei ist die von der ERA veröffentlichte Liste mit TSI-Mängeln zu berücksichtigen. Sind TSI anwendbar, wird durch den Antragssteller der zum Nachweis der TSI-Konformität des zu genehmigenden Fahrzeuges geeigneten Konformitätsnachweis ermittelt.

Neben den anwendbaren TSI sind auch die nichtanwendbaren TSI zu ermitteln und ggf. ein Antrag zur Nichtanwendung nach Art. 7 der RL (EU) 2016/797 bei der je nach Verwendungsgebiet zuständigen nationalen Sicherheitsbehörde zu stellen.

Soweit neue, anwendbare TSI Übergangsmaßnahmen vorsehen, kann der Antragsteller im Rahmen der Übergangszeit Anforderungen aus dieser neuen Richtlinie wählen. Falls der Antragsteller davon Gebrauch macht, hat er die Kohärenz der Maßnahmen darzulegen, die jeweilige TSI-Version im Genehmigungsantrag anzugeben und ggf. einen Antrag auf Änderung des „Standpunktes zur Vorbereitung“ zu stellen. Bezüglich dieser

Vorgaben besteht allerdings keine Pflicht des Antragstellers einen Antrag auf Nichtanwendung nach Art. 7 der RL (EU) 2016/797 zu stellen (Art. 17 Abs. 4 lit. c) DVO (EU) 2018/545). Entsprechend kann der Antragsteller hinsichtlich der nationalen Vorschriften der Mitgliedsstaaten verfahren, sofern diese es zulassen.

Soweit eine zu überarbeitende TSI noch nicht angenommen ist, kann der Antragsteller die Konformität durch einen EG-Konformitätsnachweis gem. Art. 6 Abs. 3 der RL (EU) 2016/797 nachweisen. Der Nachweis der Einhaltung der nationalen Vorschriften kann durch einen geeigneten Konformitätsnachweis gem. Art. 13 Abs. 2 der RL (EU) 2016/797 erfolgen.

vi) Ermittlung und Festlegung der Maßnahmen, die getroffen werden müssen, um mit dem Fahrzeug Probefahrten im Netz durchführen zu können

In Vorbereitung des Antrages ermittelt der Antragsteller gemäß Art. 18 DVO (EU) 2018/545 auch die (nationalen) Vorschriften, die die **Durchführung von Probefahren** im jeweiligen Netz regeln. Aus diesen leitet er die zur Durchführung der Probefahrten notwendigen Maßnahmen ab.

vii) Befristete Genehmigung zur Nutzung eines Fahrzeugs für Probefahrten im Netz

Art. 19 DVO (EU) 2018/545 sieht vor, dass befristete **Genehmigungen für Probefahrten** nur nach nationalen Vorschriften sowie im Einklang mit diesen durch die jeweils zuständigen NSB erteilt werden dürfen.

viii) Ermittlung der voraussichtlichen Nutzungsbedingungen für ein Fahrzeug und sonstige Beschränkungen

Weiterhin hat der Antragsteller die voraussichtlichen **Nutzungsbedingungen und Beschränkungen** für den geplanten Einsatz des Fahrzeuges/Fahrzeugtyps zu ermitteln.

ix) Ermittlung von Konformitätsbewertungen

Letztlich hat der Antragsteller nach Art. 21 der RL (EU) 2018/545 im Rahmen der Vorbereitung (preparation) des Antrages auch die Konformitätsbewertungen, die nach Anhang IV der RL (EU) 2016/797 erforderlich sind, zu ermitteln.

Es muss also ermittelt werden, **welche jeweiligen EG-Prüferklärungen** – beruhend auf dem Produktzulassungssystem des „Blue Guide“ – für das zuzulassende Teilsystem und dessen Teilsysteme **benötigt werden**.

b) Phase 2: Vorbereitung (Kapitel 3 der DVO (EU) 2018/545)

Kapitel 3 der DVO (EU) 2018/545 beschreibt die Vorbereitungen, die zu treffen sind, bevor der Antrag eingereicht werden kann.

Hierbei ist die deutsche Variante unglücklich übersetzt. Aus dem Titel des Kapitels „Vorbereitung“ wird der Unterschied zum vorausgegangenen Kapitel, in dessen Bezeichnung ja ebenfalls die Vorbereitung enthalten ist, wenig klar. Die englischen Kapiteltitel – preparation of the application (Chapter 2), sowie pre engagement (Chapter 3) – machen den Unterschied dagegen deutlicher. So ist das pre engagement als ein dem

eigentlichen Antrag vorgelagertes Abstimmungsverfahren zwischen der Genehmigungsbehörde und dem Antragssteller zu verstehen, um die Rahmenbedingungen des Antrages im Vorfeld abzuklären und eventuelle Unstimmigkeiten aus dem Weg zu räumen.

i) Vorbereitung

Im Rahmen der Vorbereitung – **pre engagement** – in der zweiten Phase des Zulassungsprozesses kann der Antragsteller an die Genehmigungsstelle und die NSB Vorbereitungsanträge stellen (Art. 22 DVO (EU) 2018/545). Diese führen zur Ausarbeitung des sog. **Standpunkt zur Vorbereitung** (in der englischen Version der Durchführungsverordnung die als „baseline" bezeichnet), der im Vorfeld der Antragsstellung die Erwartungen der Genehmigungsstelle und der NSB festsetzt. Zudem dient die Vorbereitung der **Klärung, Nachverfolgung und dem Nachvollziehen von Problemen**, die mit dem Antragsteller erörtert wurden.[5] Der Antrag auf „Vorbereitung"/Vorbereitungsantrag muss dabei förmlich über die zentrale Anlaufstelle eingereicht werden und von einem Dossier („Vorbereitungsdossier") begleitet werden, welches zumindest die in Art. 23 DVO (EU) 2018/545 beschriebenen erforderlichen Informationen beinhaltet.

Das pre engagement enthält auch eine zeitliche Komponente. Der Antragsteller hat nach Art. 22 Abs. 2 DVO (EU) 2018/545 den Genehmigungsantrag spätestens 84 Monate (7 Jahre) nach der Bekanntgabe der Stellungnahme gem. Art. 24 Abs. 2 DVO (EU) 2018/545 zu stellen.

Zu beachten ist hier, dass in Art. 22 Abs. 2 lediglich die Stellungnahme nach Art. 24 Abs. 2 als Beginn der Frist gilt. So ist eine geänderte Stellungnahme nach Art. 24 Abs. 5 für die Fristberechnung nicht maßgeblich, die Frist läuft während des Änderungsprozesses weiter.

Während des Vorbereitungsprozesses, als der zweiten Phase der Erteilung der Inbetriebnahmegenehmigung, ist der **Antragsteller an die von ihm gewählte Genehmigungsstelle gebunden**. Die Bindung erlischt erst mit Beendigung der Vorbereitung, sei es weil der Antrag auf Inbetriebnahmegenehmigung gestellt wird, die oben genannte Frist verstrichen ist oder ein Antrag des Antragstellers gestellt wird, die Vorbereitung zu beenden (Art 22. Abs. 3, Abs. 4, Abs. 5 DVO (EU) 2018/545, wobei Abs. 4, inhaltlich redundant, lediglich Abs. 3 exemplifiziert).

ii) Vorbereitungsdossier

Das mit dem Antrag zur Vorbereitung einzureichende **Vorbereitungsdossier** muss mindestens (vgl. Art. 22 Abs. 1 DVO (EU) 2018/545) die in Art. 23 DVO (EU) 2018/545 aufgeführten Inhalte enthalten:

- eine Beschreibung des zu genehmigenden Fahrzeugtyps und/oder Fahrzeugs, gegebenenfalls einschließlich der vorgesehenen Varianten und/oder Versionen, sowie eine Beschreibung der im Hinblick auf deren Entwicklung durchgeführten Aufgaben und Tätigkeiten;

[5] Auch hier ist die deutsche Version/Übersetzung des Verordnungstextes ungenau bis missverständlich. So spricht sie, ausgesprochen hypotaktisch, von der „Verfolgung" von Problemen. Hier ist der englische Text wieder eindeutiger, der von *„tracking issues"* ausgehend weniger die Komponente es Erreichens bzw. Einholens, noch die des Lösens, Umsetzens in den Vordergrund stellt, sondern vielmehr die beobachtende Bedeutung eines „Nachverfolgens" oder vielmehr „Nachvollziehens" beinhaltet.

- die vom Antragsteller gewählte Genehmigungsstelle und das bzw. die gewählte(n) Genehmigungsverfahren nach Art. 14;
- eine Beschreibung des Verwendungsgebiets;
- eine Beschreibung der voraussichtlichen Nutzungsbedingungen für das Fahrzeug und sonstigen Beschränkungen gemäß Art. 20;
- die Zeitplanung des Antragstellers für seine Aufgaben im Rahmen des Fahrzeuggenehmigungsverfahrens, gegebenenfalls einschließlich geplanter Probefahrten im Netz;
- eine Beschreibung der Methodik zur Erfassung der Anforderungen gemäß Art. 13;
- eine Aufstellung der Vorschriften und Anforderungen, die der Antragsteller nach Art. 17 und Art. 18 zu erfüllen hat;
- eine Aufstellung der gemäß Art. 21 ermittelten Konformitätsbewertungen, gegebenenfalls einschließlich der anzuwendenden Module sowie Zwischenprüfbescheinigungen;
- gegebenenfalls eine Beschreibung der praktischen Modalitäten für die Nutzung des Fahrzeugs für Probefahrten im Netz;
- eine Aufstellung des Inhalts der Unterlagen, die der Antragsteller der Genehmigungsstelle und den für das Verwendungsgebiet zuständigen NSB im Hinblick auf die Beantragung der Fahrzeugtypgenehmigung und/oder der Genehmigung für das Inverkehrbringen von Fahrzeugen vorzulegen beabsichtigt;
- einen Vorschlag gemäß Art. 10, welche Sprache im Rahmen des Fahrzeuggenehmigungsverfahrens verwendet werden soll;
- eine Beschreibung der Organisation des Antragstellers mit Blick auf seine Aufgaben im Rahmen des Fahrzeuggenehmigungsverfahrens einschließlich Kontaktangaben des Antragstellers, Informationen über Kontaktpersonen, Ersuchen um Koordinierung und Treffen mit der Genehmigungsstelle und den für das Verwendungsgebiet zuständigen NSB.

iii) Standpunkt zur Vorbereitung

Nachdem der Vorbereitungsantrag gestellt wurde muss die Genehmigungsstelle binnen eines Monats erklären, dass das Vorbereitungsdossier vollständig ist oder – sofern erforderlich – weitere Unterlagen anfordern. Anschließend wird der **Standpunkt zur Vorbereitung**, einschließlich der Festsetzung der relevanten TSI Versionen, der nationalen Vorschriften und der Angabe der zu verwendenden Sprache in der Stellungnahme der Genehmigungsstelle festgehalten (Art. 24 DVO (EU) 2018/545). Sollten sich Änderungen ergeben, die Auswirkungen auf den Standpunkt zur Vorbereitung haben, stellt der Antragsteller einen geänderten/aktualisierten Vorbereitungsantrag. Dieser umfasst dabei nur die Änderungen und die diesen entsprechenden Schnittstellen zum ursprünglichen Vorbereitungsantrag. Dabei kommen vor allem Änderungen des Baumusters oder der Bewertungsmethode in Betracht, die aus wesentlichen Sicherheitserwägungen resultieren. Aber auch rechtliche Änderungen, die den Standpunkt ungültig machen oder Änderungen, die vom Antragsteller freiwillig eingeführt wurden. Der geänderte Vorbereitungsantrag wird sodann von der Genehmigungsstelle und gegebe-

nenfalls der NSB binnen eines Monats geprüft und Stellung bezogen. Diese Stellungnahme wird in einer aktualisierten Version des Standpunktes zur Vorbereitung festgehalten.

c) Phase 3: Konformitätsbewertung (Kapitel 4 der DVO (EU) 2018/545)

In der dritten Phase des Zulassungsprozesses geht es darum, dass der Antragsteller für die **notwendigen Konformitätsbewertungen** zu sorgen hat.

i) Konformitätsbewertung

Die nach Anhang I DVO (EU) 2018/545 nachzuweisenden Konformitätsbewertungen sind von den Konformitätsbewertungsstellen durchzuführen, die obendrein für die Zusammenstellung der erforderlichen Unterlagen sowie das Erstellen der Konformitätsbewertungsberichte verantwortlich sind (Art. 25 DVO (EU) 2018/545).

ii) Überprüfungen und Nachweise

Der Antragsteller veranlasst die **erforderlichen Prüfungen** für die nach Anhang I DVO (EU) 2018/545 **notwendigen Nachweise** (Art. 26 DVO (EU) 2018/545). Dabei schreiben die Genehmigungsstelle und die nationalen Sicherheitsbehörden keine Anforderungen an die Nachweise vor, können jedoch bei begründeten Zweifeln die Durchführung weiterer Prüfungen verlangen.

iii) Behebung von Nichtkonformitäten

Soweit **Nichtkonformitäten** zwischen dem Fahrzeug und den TSI und/oder nationalen Vorschriften vorliegen, sind diese vom Antragsteller zu **beheben** (Art. 27 Abs. 1 DVO (EU) 2018/545). Dazu werden dem Antragsteller nach Art. 27 Abs. 2 zum einen die Möglichkeit der Änderung des Entwurfes in den nicht konformen Belangen sowie zum anderen die Möglichkeit der Festlegung von Nutzungsbedingungen zur konformen Verwendung gegeben. Auch hier zeigt sich wieder einmal die schlechte handwerkliche Qualität des mit Füllwörtern versehenen Gesetzestextes. Die alternative Verwendung der beiden Möglichkeiten zu Behebung (nicht Minderung) der Nichtkonformitäten ist wohl kaum als „alternativ“ zu verstehen, vielmehr können diese Werkzeuge sowohl nebeneinander als auch mehrfach („eine oder mehrere“) angewandt werden, um die Nichtkonformitäten zu beheben.

d) Phase 4: Antragstellung (Kapitel 5 der DVO (EU) 2018/545)

Auf die Konformitätsbewertung folgt die **Antragstellung**. In dieser hat der Antragsteller die zuvor zusammengestellten Unterlagen gesammelt der Bewertungsstelle einzureichen.

i) Erbringung der für den Antrag notwendigen Nachweise

Nach Art. 28 DVO (EU) 2018/545 hat der Antragsteller die notwendigen **Nachweise zu erbringen**, indem er die EG-Prüferklärungen der einzelne Teilsysteme zusammenstellt und in das vorzulegende Begleitdossier (vgl. Art. 29) einfügt. Zudem hat der Antragsteller sicherzustellen, dass nicht definierte Schnittstellen bei der Erfassung der Anforderungen berücksichtigt wurden und den grundlegenden Anforderungen nach Art. 3 Abs. 1 RL (EU) 2016/797 entsprechen.

ii) Zusammenstellung des dem Antrag beigefügten Dossiers

Die Verordnung spezifiziert in diesem Artikel noch einmal, dass der Antragsteller das dem Antrag beizufügende **Dossier strukturiert zu erstellen** hat (Art. 29 Abs. 1 DVO (EU) 2018/545). Für die Genehmigungen abseits der Erstgenehmigung ist zusätzlich noch die Gültigkeit der vorhandenen Fahrzeugtypengenehmigung zu prüfen, bei einer geplanten Erweiterung des Verwendungsgebietes und einer neuen Genehmigung bei Auf- / bzw. Umrüstung sind zu den entscheidungserheblichen Unterlagen alle dem Dossier für die frühere Genehmigung beigefügten Unterlagen vorzulegen.

iii) Inhalt und Vollständigkeit des Antrags

Der Antrag wird von der Genehmigungsstelle als vollständig angesehen, wenn er den Anforderungen des Anhang I der DVO (EU) 2018/545 entspricht (Art. 30 Abs. 1 DVO (EU) 2018/545). Zudem stellt Art. 30 noch weitere Anforderungen für Anträge auf eine Erweiterung des Verwendungsgebietes. So ist bei diesen das Dossier auf Aspekte zu beschränken, die die nationalen Vorschriften und die technische Kompatibilität zwischen Fahrzeug und Netz im erweiterten Verwendungsgebiet betreffen (Art. 30 Abs. 2 lit. a) DVO (EU) 2018/545). Soweit das Fahrzeug / der Fahrzeugtyp gemäß der RL 2008/57/EG oder früher genehmigt wurde sind zudem die entsprechenden nationalen Vorschriften hinzuzufügen (Art. 30 Abs. 2 lit. c) DVO (EU) 2018/545). Wenn die ursprüngliche Genehmigung die Nichtanwendung von TSI vorsah, hat der Antragsteller die entsprechenden Entscheidungen über die Nichtanwendung der TSI, die sich auf das erweiterte Verwendungsgebiet beziehen dem vollständigen Begleitdossier hinzuzufügen.

iv) Antragsstellung über die zentrale Anlaufstelle

Es folgt die eigentliche Antragstellung. Diese hat über die **zentrale Anlaufstelle**, den **One Stop Shop** (OSS) der ERA (Art. 12 Verordnung (EU) 2016/796) zu erfolgen (Art. 31 Abs. 1 DVO (EU) 2018/545).

Dabei hat der Antragsteller die Wahl, ob für ihn die Agentur oder die zuständige NSB als Genehmigungsstelle fungieren soll. Hat sich der Antragsteller entschieden, so kann die Genehmigungsstelle nicht mehr geändert werden, bis das Verfahren abgeschlossen ist. Der OSS leitet das Dossier des Antragstellers an die gewählte Genehmigungsstelle weiter.

e) Phase 5: Bearbeitung des Antrages (Kapitel 6 der DVO (EU) 2018/545)

Das sechste Kapitel der Durchführungsverordnung regelt die **Bearbeitung des Antrages durch die gewählte Genehmigungsstelle** in den Art. 32 bis 46 DVO (EU) 2018/545.

i) Vollständigkeitsprüfung

Zu Beginn der Bearbeitungsphase prüft die Genehmigungsstelle zunächst die **Vollständigkeit der vorgelegten Dokumente** (Art. 32 Abs. 1 DVO (EU) 2018/545). Die zuständigen NSB prüfen, ob das **Verwendungsgebiet konkret spezifiziert** ist und weisen auf Probleme im Zusammenhang mit der Vollständigkeit derjenigen Dokumente hin, die für die Bewertung der anwendbaren nationalen Vorschriften vorgelegt wurden. Neben der Prüfung der Vollständigkeit der eingereichten Unterlagen prüfen die Genehmigungsstelle und NSB auch deren **Geeignetheit zur Bewertung** nach den Art. 38 bis 40 DVO (EU) 2018/545).

ii) Eingangsbestätigung

Der **Eingang** des Antrages wird automatisch **durch den OSS bestätigt**. Am Tage des Eingangs beginnt auch die Bewertung des Antrages (Art. 33 DVO (EU) 2018/545).

iii) Fristen für die Bewertung des Antrags

Binnen eines Monats nach Eingang der Unterlagen hat die Genehmigungsstelle die Vollständigkeit der Unterlagen zu prüfen und das Ergebnis dieser Prüfung dem Antragsteller mitzuteilen. Von diesem Zeitpunkt an, hat die Genehmigungsstelle vier Monate Zeit über den Antrag zu entscheiden (Art. 34 Abs. 1 DVO (EU) 2018/545).

Soweit es sich um einen Antrag nach Art. 14 Abs. 1 lit. e) der DVO (EU) 2018/545 handelt, muss die Genehmigungsstelle binnen eines Monats über den Antrag entscheiden (Art. 34 Abs. 3).

Soweit der Antrag nicht vollständig war und dies dem Antragsteller mitgeteilt wurde, entscheidet die Genehmigungsstelle binnen vier Monaten nach Vorlage der fehlenden Informationen. Sofern der Antrag dann noch andere Mängel aufweist wird er abgelehnt.

Auch bei vollständigen Anträgen kann die Genehmigungsstelle oder zuständige NSB im Laufe des Verfahrens zusätzliche Informationen anfordern. Für deren Vorlage hat sie eine angemessene Frist zu setzen.

Die Genehmigungsstelle oder zuständige NSB kann die Bewertung aussetzen und die Frist aus Art. 21 Abs. 4 der RL (EU) 2016/797 verlängern, wenn begründete Zweifel bestehen und/oder der Antragsteller weitere Informationen vorzulegen hat. Eine solche Fristverlängerung ist in einer ordnungsgemäß dokumentierten Vereinbarung zwischen Antragssteller und der jeweiligen Stelle festzuhalten. Weiterhin muss die Frist in einem angemessenen Verhältnis zur Möglichkeit der Informationsbeschaffung stehen. Nach Vorlage der Informationen wird die Bewertung fortgeführt.

iv) Kommunikation während des Antrags

Sämtliche Kommunikation hinsichtlich der Einstufung von Problemen bzw. den in Art. 41 DVO (EU) 2018/545 genannten Belangen findet **über den OSS** statt. Ebenso werden alle Sachstandsabfragen über diesen abgewickelt. Vereinbarungen über die Kommunikation zwischen den Beteiligten stellen die ERA sowie die NSB in ihren jeweiligen Leitlinien auf.

v) Informationsmanagement während des Antrags

Nach Art. 36 DVO (EU) 2018/545 sind sämtliche Zwischenergebnisse und Endergebnisse der Antragsbearbeitung **in den OSS** einzupflegen. Soweit die NSB ein anderes Informationsmanagementsystem verwenden, übermitteln sie alle einschlägigen Informationen an die zentrale Anlaufstelle.

vi) Koordinierung der Antragsbewertung zwischen der Genehmigungsstelle und den für das Verwendungsgebiet zuständigen NSB

Die **Genehmigungsstelle koordiniert sich mit der NSB** (Art. 37 DVO (EU) 2018/545). So hat sie auch mit der NSB die jeweiligen Ergebnisse der Prüfung zu erörtern, bevor die

Entscheidung über den Antrag getroffen wird. Über die Koordinationstätigkeiten werden bei der zentralen Anlaufstelle Aufzeichnungen geführt. Über diese werden auch dem Antragsteller die Gründe über die Entscheidung mitgeteilt.

vii) Bewertung des Antrags

Die Genehmigungsstelle und die zuständige NSB **bewerten den Antrag**, um hinreichend sicherzustellen, dass die rechtlichen und technischen Anforderungen an das Fahrzeug erfüllt und überprüft worden sind (Art. 38 DVO (EU) 2018/545).

viii) Antragsbewertung durch die Genehmigungsstelle (Art. 39)

Die Genehmigungsstelle bewertet die in Anhang II DVO (EU) 2018/545 genannten Aspekte. Wenn eine Fahrzeugtypengenehmigung und/oder eine Genehmigung für das Inverkehrbringen für ein bestimmtes Verwendungsgebiet erteilt werden soll und der Antragsteller die zuständige NSB als Genehmigungsstelle gewählt hat, führt diese auch die Bewertung nach Anhang III der DVO (EU) 2018/545 durch.

Die Genehmigungsstelle prüft die Vollständigkeit, Relevanz und Kohärenz der Nachweise, unabhängig vom jeweiligen Verfahren. Bei neuen Genehmigungen nach Art. 14 Abs. 1 lit. d) DVO (EU) 2018/545 beschränkt sich die Prüfung auf die geänderten Komponenten sowie deren Auswirkungen auf die bestehenden Fahrzeugteile. Für Genehmigungen zur Erweiterung des Verwendungsgebietes beschränken sich die durchzuführenden Prüfungen auf die anwendbaren nationalen Vorschriften und die Kompatibilität zwischen Fahrzeug und dem Netz des erweiterten Verwendungsgebiets.

Die Genehmigungsstelle erstellt ein **Bewertungsprotokoll**, das eine eindeutige Erklärung über das Ergebnis der Bewertung in Bezug auf das Nutzungsgebiet und eventuelle Nutzungsbedingungen oder -beschränkungen enthält. Darüber hinaus umfasst das Bewertungsprotokoll eine Zusammenfassung der durchgeführten Bewertungen, einen Bericht auf Grundlage des Problemprotokolls für das betreffende Verwendungsgebiet sowie eine ausgefüllte Checkliste mit Nachweisen, dass alle in Anhang II und ggfs. in Anhang III genannten Aspekte geprüft wurden.

ix) Antragsbewertung durch die für das Verwendungsgebiet zuständigen NSB

Die Antragsbewertung durch die zuständige NSB (Art. 40 DVO (EU) 2018/545) entspricht im Wesentlichen der Prüfung seitens der Genehmigungsstelle nach Art. 39. Ihr Prüfungsumfang bezieht sich auf die in Anhang III der DVO (EU) 2018/545 genannten Aspekte. Auch das Bewertungsdossier entspricht dem nach Art. 39, allerdings mit Inhalten zu Anhang III.

x) Einstufung von (erkannten) Problemen

Die Genehmigungsstelle (und ggfs. die zuständige NSB) müssen nach Art. 41 DVO (EU) 2018/545 die während der Bewertung festgestellten Probleme im Bewertungsprotokoll nach verschiedenen Kategorien getrennt festhalten:

- Kategorie 1: Probleme, die für das Verständnis des Antragsdossiers eine Antwort des Antragstellers erfordern.
- Kategorie 2: Probleme, die eine Änderung des Antragsdossiers oder eine geringfügige Maßnahme des Antragsstellers nach sich ziehen können, wobei

die zu ergreifende Maßnahme im Ermessen des Antragsstellers und der Erteilung der Genehmigung nicht im Weg steht.

- Kategorie 3: Probleme, die eine Änderung des Antragsdossiers durch den Antragsteller erfordern, aber dem Erteilen der Genehmigung unter zusätzlichen und/oder restriktiveren Nutzungsbedingungen und anderen als den angegebenen Beschränkungen, nicht im Weg stehen. Das Problem muss für die Erteilung der Genehmigung gelöst sein, die Korrekturmaßnahmen sind vom Antragsteller vorzuschlagen und bedürfen vorheriger Absprachen zwischen den Beteiligten Parteien.
- Kategorie 4: Probleme, die eine Änderung des Antragsdossiers durch den Antragsteller erfordern. Die Korrekturmaßnahmen sind durch den Antragsteller vorzuschlagen und entsprechend abzustimmen. Probleme dieser Kategorie 4 sind insbesondere auch Nichtkonformitäten im Sinne des Art. 26 Abs. 2 der RL (EU) 2016/797. Die Genehmigung darf nur erteilt werden, wenn das Problem gelöst wurde.

Nach einer Antwort oder dem Ergreifen einer beseitigenden Maßnahme durch den Antragsteller nehmen die Genehmigungsstelle oder NSB eine Neubewertung des Problems vor und stufen es anschließend als „offenes“ oder „abgeschlossenes“ Problem ein.

xi) Begründete Zweifel

Nach Art. 2 Nr. 7 der DVO (EU) 2018/545 versteht man unter einem **begründeten Zweifel** „ein von der Genehmigungsstelle und/oder den für das Verwendungsgebiet zuständigen NSB festgestelltes Problem der Kategorie 4 gemäß Artikel 41 Abs. 1 lit. d) DVO (EU) 2018/545 hinsichtlich der vom Antragsteller im Antrag gemachten Angaben mit einer Begründung und entsprechenden Nachweisen“.

Liegen begründete Zweifel vor, können die Genehmigungsstelle und/oder die NSB verschiedene **Maßnahmen (auch kumulativ) ergreifen** (Art. 42 Abs. 1 DVO (EU) 2018/545). So können die im Antrag vorgebrachten Informationen einer detaillierteren Prüfung unterzogen, vom Antragsteller zusätzliche Informationen angefordert oder vom Antragsteller die Durchführung von Probefahrten im relevanten Netz verlangt werden.

In der Aufforderung der Genehmigungsstelle oder NSB an den Antragsteller muss angegeben werden, in welcher Sache der Antragsteller tätig werden muss, jedoch nicht Art oder Inhalt der Korrekturmaßnahmen. Wie der Antragsteller der Aufforderung nachkommt liegt in dessen Ermessen. Bezüglicher der vom Antragsteller vorgeschlagenen Maßnahmen zur Behebung des Problems stimmen sich die Genehmigungsstelle und die NSB untereinander ab.

Für die Bearbeitung der begründeten Zweifel ist das **Problemprotokoll** nach Art. 41 DVO (EU) 2018/545 zu verwenden, die lit a) bis c) des Art. 42 Abs. 4 sind entsprechend redundant, im Fall des lit. a) liegt sogar ein Zirkelschluss vor, da gemäß der Definition des begründeten Zweifels nur ein solches Problem erfasst ist, das auch in der Kategorie 4 nach Art. 41 Abs. 1 lit. d) fällt.

Soweit sich der Antragsteller bereit erklärt auf Verlangen der Genehmigungsstelle zusätzliche Informationen vorzulegen wird die Frist für deren Bereitstellung nach Art. 34

Abs. 5, 6 DVO (EU) 2018/545 festgelegt. Ist der Antragsteller dazu nicht bereit, entscheidet die Genehmigungsstelle auf Grund der verfügbaren Informationen.

Wenn ein begründeter Zweifel durch die Verwendung von Nutzungseinschränkungen oder (restriktiveren) -bedingungen behoben werden kann und der Antragsteller diesen zustimmt, kann eine Genehmigung unter diesen erteilt werden. Hier ist allerdings das Zustimmungserfordernis schon fraglich, da diese eigentlich durch den Antragsteller vorgeschlagen werden müssen. Insofern wäre dieser Punkt des Art. 42 Abs. 6 ebenfalls entbehrlich. Er scheint (wie auch andere Stellen der Durchführungsverordnung) vielmehr der Verdeutlichung (im Sinne einer Anleitung) als der konkreten Regelung zu dienen, was an sich begrüßenswert und Bürgernah sein mag, jedoch im falschen Text seinen Niederschlag findet.

xii) Von der Genehmigungsstelle durchzuführende Prüfungen bezüglich der Bewertungen der für das Verwendungsgebiet zuständigen NSB

Die **Genehmigungsstelle prüft**, ob die Bewertungen der zuständigen NSB und das **Ergebnis der Prüfung** nach Art. 40 Abs. 6 lit. a) DVO (EU) 2018/545 miteinander in Einklang stehen. Sofern dies der Fall ist, überprüft sie, ob die **Checklisten** nach Art. 40 Abs. 6 lit. d) **vollständig ausgefüllt** und alle relevanten **Probleme abgeschlossen** wurden (Art. 43 Abs. 1 und 2 DVO (EU) 2018/545).

Soweit die Bewertungen nicht im Einklang stehen, fordert die Genehmigungsstelle die NSB auf, die Gründe dafür zu untersuchen, woraufhin die Genehmigungsstelle ihre Bewertung gem. Art 39 überprüft (Art. 43 Abs. 3 lit. a)) und/oder die NSB ihre Bewertung überprüft (Art. 43 Abs. 3 lit. b)). Die Ergebnisse der Untersuchung werden allen NSB mitgeteilt, die mit dem Antrag auf die Erteilung der Genehmigung befasst sind.

Falls eine der Checklisten noch unvollständig ist oder noch nicht abgeschlossene Probleme bestehen, wird die zuständige NSB durch die Genehmigungsstelle aufgefordert diese zu untersuchen. Die zuständigen NSB beantworten die Fragen der Genehmigungsstelle im Hinblick auf unvollständige Checklisten, nicht abgeschlossene Probleme und Unstimmigkeiten zwischen den Bewertungen. Die Genehmigungsstelle trägt den Bewertungen der anwendbaren nationalen Vorschriften durch die NSB vollständig Rechnung. Dabei beschränkt sich der Umfang der Prüfung der Genehmigungsstelle auf die Übereinstimmung und Vollständigkeit der Bewertungen. Bei Differenzen zwischen Genehmigungsstelle und zuständiger NSB wird das Schiedsverfahren nach Art. 21 Abs. 7 der RL (EU) 2016/797 bemüht.

xiii) Schiedsverfahren

Falls die Agentur als Genehmigungsstelle tätig wird, kann diese das Genehmigungsverfahren aussetzen (Art. 44 DVO (EU) 2018/545). Dies jedoch nur in Abstimmung mit den am Verfahren beteiligten zuständigen NSB und nur für die Dauer der Zusammenarbeit, die zur Erreichung einer für alle Seiten annehmbaren Bewertung notwendig ist, und ggfs. bis die Beschwerdekammer in der in Art. 21 Abs. 7 RL (EU) 2016/797 eine Entscheidung trifft. Dem Antragsteller werden die Aussetzungsgründe durch die Agentur mitgeteilt.

xiv) Ergebnis der Antragsbewertung

Die Genehmigungsstelle stellt die ordnungsgemäße Durchführung der Antragsbewertung sicher, indem sie prüft, ob

- die einzelnen Stufen korrekt durchgeführt wurden;
- hinreichende Nachweise vorliegen, dass alle relevanten Aspekte des Antrags untersucht wurden
- bzgl. Problemen der Kategorien drei und vier sowie zu Anforderungen zusätzlicher Informationen schriftliche Antworten des Antragstellers vorliegen;
- alle Probleme der Kategorien drei und vier gelöst wurden, oder falls diese nicht gelöst wurden dies klar begründet wurde;
- die Bewertungen der Entscheidungen dokumentiert, fair und kohärent sind;
- die Feststellungen auf den Bewertungsdossiers basieren und die Bewertung als Ganzes widerspiegeln.

Wenn aus ihrer Sicht die Antragsbewertung ordnungsgemäß durchgeführt wurde, genügt eine – ggfs. mit Bemerkungen (Art. 45 Abs. 2) – versehene **schriftliche Bestätigung der ordnungsgemäßen Anwendung** des Art. 45 Abs. 1 DVO (EU) 2018/545. Ist dies nicht der Fall, so müssen die Gründe für die Feststellung einer nicht ordnungsgemäßen Antragsbewertung spezifisch festgehalten werden (Art. 45 Abs. 3).

Abschließend erstellt die Genehmigungsstelle ein **Bewertungsdossier inklusive einer Begründung** wie sie zu dessen Ergebnis gelangt ist. Das Bewertungsdossier enthält die Entscheidungen hinsichtlich der ordnungsgemäßen Antragsbewertung oder nicht ordnungsgemäßen Antragsbewertung nach diesem Artikel. Diesem werden die Bewertungsdossiers der Bewertung nach Art. 39 Abs. 5 und 40 Abs. 5 der DVO (EU) 2018/545 zugrunde gelegt.

xv) Entscheidung zur Genehmigung oder Ablehnung des Antrags

Die **Entscheidung der Genehmigungsstelle** über den Antrag erfolgt unbeschadet Art. 34 DVO (EU) 2018/545 **binnen einer Woche nach Abschluss der Bewertung**. Dabei wird die Genehmigung erteilt, wenn die Bewertung der Aspekte aus den Anhängen II und III hinreichende Gewähr bieten, dass der Antragsteller und die ihn unterstützenden Akteure im Einklang mit Art. 38 ihre Aufgaben im erforderlichen Umfang erfüllen. Andernfalls wird die Genehmigung nicht erteilt.

In der Genehmigung gibt die Bewertungsstelle die etwaigen Nutzungsbedingungen und sonstigen Beschränkungen, die Gründe für die Entscheidung und die Möglichkeiten und Mittel Beschwerde gegen die Entscheidung inklusive der entsprechenden Fristen an.

Nutzungsentscheidungen müssen in Einklang mit den grundlegenden Konstruktionsmerkmalen des Fahrzeugs stehen und dürfen nicht befristet oder beschränkt erlassen werden, es sei denn, dass die Konformität mit TSI und/oder nationalen Vorschriften nicht vor der Erteilung nachgewiesen werden konnte und dies daher notwendig ist und/oder die TSI oder nationalen Vorschriften eine Schätzung verlangen, wann Konformität hergestellt ist.

Wird der Antrag abgelehnt oder an andere Nutzungsbedingungen bzw. Einschränkungen geknüpft als beantragt, kann der **Antragsteller eine Überprüfung** nach Art. 51 DVO (EU) 2018/545 **beantragen**. Sollte er mit der Antwort der Genehmigungsstelle nicht zufrieden sein, so kann er bei der zuständigen Behörde ein Schiedsverfahren nach Art. 21 Abs. 7 RL (EU) 2016/797 anstreben.

Die **endgültige Entscheidung** wird **beim OSS registriert** und dem Antragsteller sowie den zuständigen NSB, mitsamt dem Bewertungsdossiers, durch diesen übermittelt.

f) Endgültige Unterlagen (Kapitel 7 der DVO (EU) 2018/545)

Das Kapitel 7 der DVO (EU) 2018/545 beschreibt die **Zusammenstellung der Unterlagen**, die durch die Genehmigungsstelle final zusammengeführt werden. Dabei stellt Art. 47 der DVO 2018/545 klar, dass die Genehmigung die in den Art. 48 und/oder Art. 49 genannten Informationen enthalten und ihr eine **Europäische Identifikationsnummer** (EIN) zugewiesen werden muss. In diesem Kapitel sind weiterhin die Regelungen zur **Registrierung der Genehmigung in ERADIS[6] und ERATV**, sowie die Anforderungen für die Überprüfung nach Art. 21 Abs. 11 der RL (EU) 2016/797 und die Archivierung des Dossiers enthalten. Für das Zulassungsverfahren im engen Sinne spiele diese jedoch eine untergeordnete Rolle.

g) Aussetzung, Änderung oder Widerruf einer erteilten Genehmigung (Kapitel 8 der DVO (EU) 2018/545)

Das letzte in der DVO enthaltene inhaltliche Kapitel regelt **Aussetzung, Änderung oder Widerruf einer erteilten Genehmigung** (Art. 53, 54 DVO (EU) 2018/545). Dabei ist eine Aussetzung als vorrübergehende Sicherheitsmaßnahme möglich, der Widerruf oder die Änderung kommen nach einer Aussetzung bei schwerwiegenden Sicherheitsverstößen in Betracht.

3) Zusammenfassender Überblick

Zusammenfassend lässt sich festhalten, dass das Fahrzeugzulassungsverfahren grundsätzlich zunächst bei der ERA angesiedelt ist, soweit es sich dabei jedoch um lediglich Projekte mit nationaler Reichweite handelt auch die nationalen Sicherheitsbehörden (NSB) die Rolle der Genehmigungsstelle erfüllen können. Diese Behörden sind zudem für die Prüfung zuständig, ob die lediglich nationalen Regelungen eingehalten wurden. Geprüft wird letztlich, ob die eingereichten Dokumente schlüssig und vollständig sind. Eine Prüfung des tatsächlichen Gerätes bleibt den bestimmten sowie den benannten Stellen vorbehalten.

[6] European Railway Agency Database of Interoperability and Safety (ERADIS).

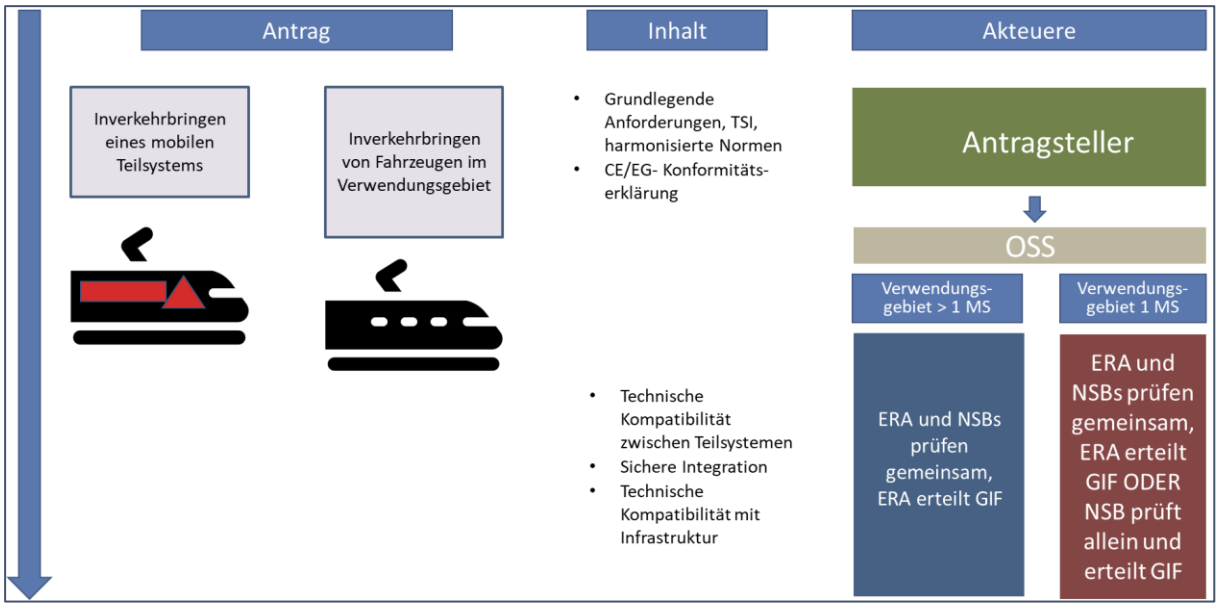


Abb. 1: Überblick Erteilung der Inbetriebnahmegenehmigung Fahrzeug (GIF)

Die eingereichten Unterlagen werden anschließend zusammen mit dem Begleitdossier – d.h. allen Informationen zum Verfahren – archiviert.

(ii) Sonstiges Teilsystem

Für die **Teilsysteme, die nicht mobil sind**, gelten andere Voraussetzungen. Für **KI-bezogene Anwendungen** ist insbesondere das strukturelle Teilsystem der streckenseitgen Zugsteuerung, Zugsicherung und Signalgebung (ZZS, vgl. Anhang II, RL (EU) 2016/797) relevant.

Sowohl die erstmalige Inbetriebnahme als auch die Aufrüstung oder Erneuerung der Teilsysteme Infrastruktur, streckenseitige Zugsteuerung Zugsicherung und Signalgebung bedarf grundsätzlich einer Inbetriebnahmegenehmigung § 9 Abs. 2, 4 Nr. 1 EIGV.

1) Zuständigkeiten

Die Zuständigkeit für die Zulassung der nicht mobilen Teilsysteme liegt gem. § 10 Abs. 2 EIGVO beim Eisenbahnbundesamt.

2) Verfahren

Im Rahmen der Erteilung der Inbetriebnahmegenehmigung für die streckenseitigen Teilsysteme Infrastruktur, Energie streckenseitige Zugsteuerung, Zugsicherung und Signalgebung sowie für die übrige Eisenbahninfrastruktur ist zunächst zu unterscheiden, ob technische Spezifikationen für die Interoperabilität anzuwenden sind, oder gerade nicht (§ 16 Abs. 1 EIGV).

a) Nichtanwendbarkeit der TSI

Soweit keine TSI Anwendung finden richten sich die Antragsvoraussetzungen nach § 17 EIGV. Nach diesem sind für strukturelle Teilsysteme: § 16 Abs. 1 S. 2 und 3 Nr. 1 lit. b), Nr. 2 bis 7 und S. 4 sowie Abs. 3 und 4 EIGV, für die übrige Eisenbahninfrastruktur: § 16 Abs. 1 S. 2 und 3 Nr. 2 bis 7 sowie Abs. 3 und 4 EIGV zu erfüllen. Diese korrespondieren mit den oben aufgeführten Nummern für Teilsysteme mit anwendbaren TSI.

b) Anwendbarkeit der TSI

Sofern die TSI anwendbar sind, ist nachzuweisen, dass die grundlegenden Anforderungen erfüllt sind. Dieser Nachweis gilt als erbracht, wenn die **folgenden Unterlagen vorgelegt** wurden:

- Die EG Prüferklärungen nach Art. 15 der RL (EU) 2016/797;
- eine Erklärung des Antragstellers, dass der Bestandteil des Eisenbahnsystems die grundlegenden Anforderungen erfüllt und insbesondere die technische Kompatibilität sowie die sichere Integration gewährleistet sind;
- einer Erklärung, dass alle ermittelten Gefährdungen auf einem vertretbaren Niveau gehalten werden und eine Bewertungsstelle einen Sicherheitsbewertungsbericht nach Durchführungsverordnung (EU) Nr. 402/2013 erstellt hat;
- eine Freigabe der geprüften Planung;
- eine Bestätigung der Verwendbarkeit der Bauprodukten und deren Teilprodukten;
- ein Nachweis über die durchgeführte Bauüberwachung sowie
- ein Nachweis der notwendigen Abnahmeprüfung.

Neben der Vorlage dieser Dokumente ist die **Zustimmung der Agentur** zu dem Vorhaben vorzulegen, wenn es die Ausrüstung mit dem Europäischen Zugsicherungs- und Zugsteuerungssystem oder dem globalen Mobilfunksystem für Eisenbahnen umfasst (§ 16 Abs. 2 EIGV).

Soweit der Antragsteller bestätigt, dass die Änderung nicht signifikant ist, hat er über die Änderungen Aufzeichnung zu führen.

Die Einhaltung der technischen Vorschriften ist dabei zu gewährleisten und durch Prüfsachverständige nachzuweisen.

c) Antrag

Im Rahmen der **Antragstellung** ist § 18 EIGV zu beachten. Das Eisenbahninfrastrukturunternehmen hat den Antrag und die zur Prüfung des Antrags erforderlichen **Unterlagen** nach § 16 Abs. 1 S. 3, Abs. 2 EIGV und nach Anlage VI der Genehmigungsstelle 24 Monate vor dem geplanten Inbetriebnahmetermin, spätestens zehn Wochen vor Baubeginn **vorzulegen** (§ 18 Abs. 1 EIGV).

Zudem hat der Antragsteller der Genehmigungsstelle zusätzlich zu dem Antrag auf Erteilung der Inbetriebnahmegenehmigung eine **Liste** der nach § 6 EIGV **anzuwendenden Vorschriften vorzulegen** (§ 18 Abs. 3 EIGV). In diese Liste sind aufzunehmen und zu begründen etwaige Abweichungen von

- den technischen Spezifikationen für die Interoperabilität,
- den entsprechenden notifizierten technischen Vorschriften und,
- soweit erforderlich, den technischen Vorschriften.

Gleichzeitig sind die stattdessen anzuwendenden Vorschriften anzugeben oder Nachweise zu führen, dass mindestens die gleiche Sicherheit gewährleistet ist.

Der Antragsteller hat weiterhin einen **Inbetriebnahmeverantwortlichen** oder einen anderen geeigneten Mitarbeiter **zu bestellen**, der insbesondere prüft und bestätigt, dass

- sicher gebaut, insbesondere die Bauüberwachung durchgeführt worden ist,
- alle notwendigen Prüfungen zur Einhaltung der grundlegenden Anforderungen einschließlich notwendiger Schnittstellenbetrachtungen durchgeführt worden sind,
- die Anforderungen und Nachweise nach § 16 Abs. 1 S. 3 Nr. 4 bis 7 EIGV vollständig erbracht worden sind,
- soweit einschlägig, alle Auflagen aus den Nachweisen nach Nr. 3 umgesetzt worden sind und
- Auflagen und Nebenbestimmungen aus Inbetriebnahmegenehmigungen beachtet sowie vorhandene Mängel innerhalb einer durch ihn zu bestimmenden, angemessenen Frist beseitigt worden sind.

Soweit von technischen Vorschriften abgewichen wird, sind **Nachweise** zu führen, dass mindestens die gleiche Sicherheit gewährleistet ist. Zu diesem Zweck ist ein Risikomanagementverfahren nach der Durchführungsverordnung (EU) Nr. 402/2013 durchzuführen. Wenn keine signifikanten Änderungen nach Artikel 4 der Durchführungsverordnung (EU) Nr. 402/2013 vorliegen, ist die Anwendung einer eigenen Sicherheitsmethode notwendig. Das Eisenbahn-Bundesamt kann auf Basis der Ergebnisse des Risikomanagementverfahrens eine Zustimmung im Einzelfall erteilen.

d) Verfahren

Die **Genehmigungsstelle prüft** die Antragsunterlagen nach Eingang **auf Vollständigkeit** und bescheinigt diese dem Antragssteller binnen eines Monats. Anschließend prüft sie die Unterlagen auf Vollständigkeit **und Nachvollziehbarkeit** und entscheidet innerhalb von vier Monaten (§ 19 Abs. 1 EIGV).

Soweit die Inbetriebnahmegenehmigung die Ausrüstung mit GSMR oder dem europäischen Zugsicherungs- und Zugsteuerungssystem betrifft, überprüft die Antragstelle zusätzlich zur Nachvollziehbarkeit der Antragsunterlagen, ob diese **Unterlagen mit der Zustimmung der Agentur** nach Artikel 19 der RL (EU) 2016/797 **übereinstimmen**.

Hat die Genehmigungsstelle begründete Zweifel an der Erfüllung der grundlegenden Anforderungen, kann sie vor der Entscheidung über die Erteilung der Inbetriebnahmegenehmigung verlangen, dass der Antragsteller **ergänzende Prüfungen** durchführen lässt und das Ergebnis dieser Prüfungen vorlegt. Wenn begründete Zweifel zur EG-Prüferklärung nach § 16 Abs. 1 S. 3 Nr. 1 lit. a) EIGV vorliegen, unterrichtet die Genehmigungsstelle die Kommission unverzüglich unter Angabe der Gründe nach Art. 16 Abs. 3 der RL (EU) 2016/797, welche ergänzenden Prüfungen durchzuführen sind.

Dabei liegen begründete Zweifel insbesondere dann vor, wenn vor der Erteilung der Inbetriebnahmegenehmigung bekannt ist, dass bei dem zu genehmigenden Bestandteil des Eisenbahnsystems oder bei einem Bestandteil des Eisenbahnsystems, der mit dem zu genehmigenden hinsichtlich der Bauweise und Funktion vergleichbar ist, die

Voraussetzungen vorliegen, unter denen die zuständige Aufsichtsbehörde Maßnahmen nach § 5a Abs. 2 S. 1 des Allgemeinen Eisenbahngesetzes treffen kann. Ebenso wenn Erkenntnisse vorliegen über die mangelhafte Aufgabenwahrnehmung

- durch benannte oder bestimmte Stellen, und diese Erkenntnisse eine Rücknahme nach § 36 Abs. 1 EIGV oder einen Widerruf nach § 36 Abs. 2 EIGV rechtfertigen können, oder
- durch Bewertungsstellen, und diese Erkenntnisse Maßnahmen nach Art. 11 Abs. 2 der Durchführungsverordnung (EU) Nr. 402/2013 rechtfertigen können.

(c) Zertifizierung einzelner Teilsysteme

Zum Erhalt der Zulassung müssen alle EG-Prüferklärungen des Teilsystems eingereicht werden. Zum Erhalt dieser Prüferklärungen werden die Teilsysteme und deren jeweilige Bestandteile von bestimmten und den benannten Stellen auf ihre Sicherheit und Funktionsfähigkeit hin überprüft.

Dabei orientieren sich diese Überprüfungen an den jeweils einschlägigen technischen Spezifikationen für Interoperabilität (TSI) sowie den anwendbaren technischen Normen (vgl. Anhang IV zur RL (EU) 2016/797).

Für die Zulassung der Teilsysteme müssen diese den grundlegenden Anforderungen gem. Anhang III zur RL (EU) 2016/797 entsprechen. Dies wird im Rahmen des Zulassungsprozess durch die Konformitätsbewertung und Sicherheitsnachweisführung nachgewiesen. Im Ergebnis heißt dies insbesondere, dass, sowohl in der Systementwicklung als auch bei der Sicherheitsnachweisführung, die einschlägigen Normen eingehalten werden sollten.

(i) Systementwicklung nach EN 50126

Dabei gilt für die Entwicklung von sicherheitsrelevanten (KI-)Anwendungen der in verschiedene Phasen aufgeteilte **Systementwicklungsprozess nach EN 50126 und EN 50129**. Im Rahmen der **Softwareentwicklung sind zudem zusätzlich noch die Anforderungen der EN 50128 zu beachten**. Jedoch berücksichtigt keine dieser Normen bislang den Einsatz von Systemen künstlicher Intelligenz.

Der RAMS[7]-Prozess- bzw. Lebenszyklus, als die Folge von Phasen und den jeweiligen Aktivitäten über die gesamte Lebensdauer eines Systems, vom Konzept bis zur Außerbetriebnahmen nach EN 50126 stellt sich wie folgt dar:

[7] RAMS ist das Akronym für Reliability, Availability, Maintainability, Safety.

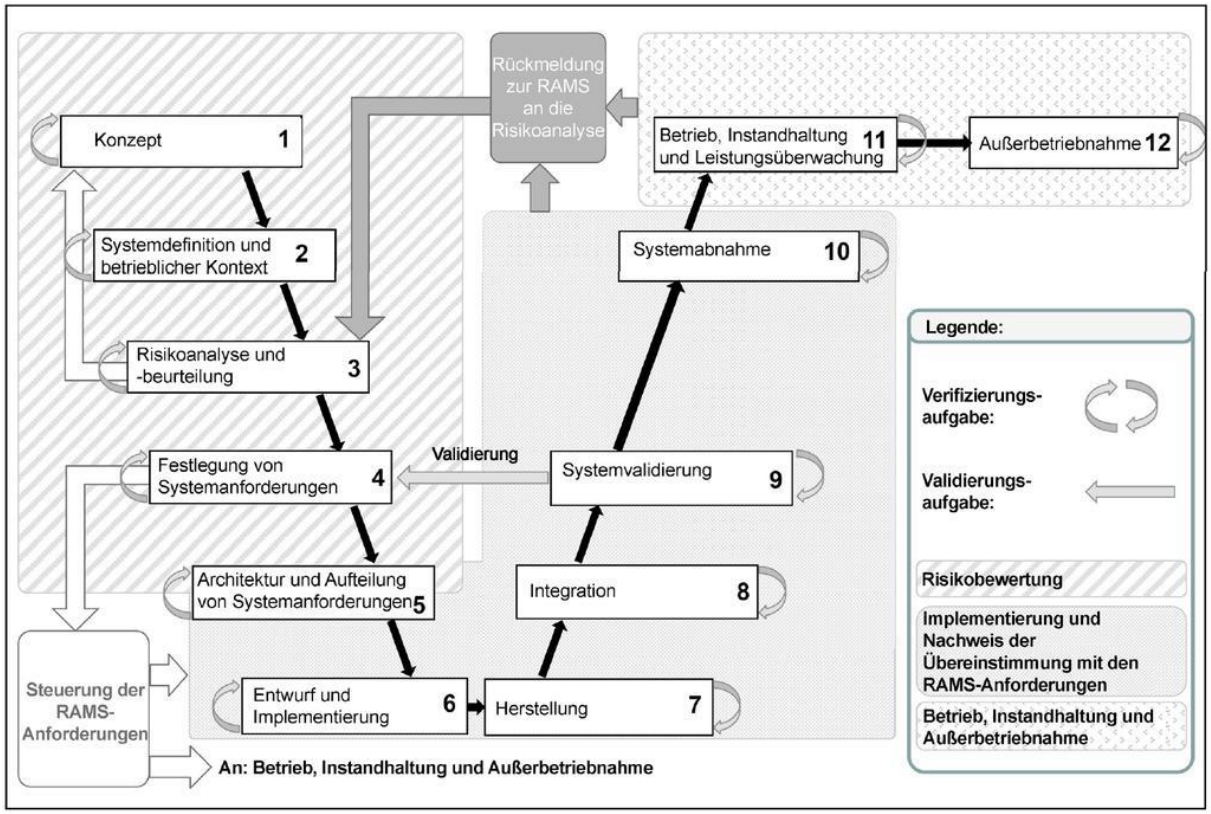


Abb. 2: Lebenszyklus V-Modell (DIN EN 50126)

1) Systemkonzept

In der Phase **Systemkonzept** werden Strategie und Ziele der Entwicklung des Systems festgesetzt. Dazu werden der Anwendungsbereich des Systems sowie dessen Kontext und Zweck definiert. Einen wichtigen Punkt stellt dabei eine **Analyse und Beschreibung der geplanten Systemumgebung** dar. Diese beinhaltet neben der technisch-physikalischen Umgebung des Systems im geplanten Einsatzbereich auch die geografischen und klimatischen Umgebungsbedingungen, die betrieblichen Randbedingungen, die Definition möglicher Schnittstellen (u.a. zu anderen Systemen) sowie der zu beachtenden Gesetze, Normen und Rahmenbedingungen.

Daneben finden in dieser Phase eine **Gegenüberstellung mit vergleichbaren Systemen** sowie eine Analyse deren Anforderungen, technischen Parameter und Sicherheitszielen statt um deren Anwendbar- oder Übertragbarkeit auf das neue System zu prüfen.

2) Systemdefinition

In der Phase der **Systemdefinition** erfolgt eine **Beschreibung des künftigen Systems einschließlich dessen Funktionalität und einer Architekturskizze**, aus der die Komponenten des Systems und die Verteilung der Aufgaben auf diese Komponenten hervorgehen. Die Anwendung des Systems im Betrieb inklusive einzelner Betriebsverfahren, Betriebsarten, Use-Cases sowie die Instandhaltungsstrategie werden in diesem Rahmen erläutert. Daneben erfolgen eine Spezifikation der Einsatzbedingungen (hinsichtlich der klimatischen, mechanischen sowie elektrischen Umgebung) sowie eine Beschreibung der Systemgrenzen und der Schnittstellen zu anderen Systemen und zum Menschen (Bediener, Instandhalter, Nutzer u.a.).

3) Risikoanalyse und -beurteilung

Im Rahmen der Phase **Risikoanalyse und -beurteilung** werden **unerwünschte Ereignisse während des Betriebs des Systems und** die daraus resultierenden **Gefährdungen ermittelt**. Hierzu werden zunächst alle Fehler identifiziert, die Ursache derartiger Ereignisse sein könnten und die Fehlerfolgen hinsichtlich ihrer Kritikalität bewertet. Zu jeder Fehlerfolge wird das akzeptable Risiko ermittelt. Ist das tatsächliche Risiko höher als der akzeptierte Wert, werden Maßnahmen zur Risikoreduktion festgelegt, um sicherzustellen, dass das Risiko auf das akzeptable Maß reduziert wird.

Das akzeptierte Risiko wird als **SIL** (safety integrity level) und **THR** (tolerable hazard rate) angegeben. Der ermittelte SIL bestimmt den Umfang der Maßnahmen zur Vermeidung von systematischen Fehlern im Entwicklungsprozess (z.B. Dokumentation, Analysen, Tests, Verifikation und Validierung). Die THR legt die zulässige Rate gefährlicher Ausfälle fest und bestimmt somit die technische Gestaltung des Systems (z.B. Bauelementeauswahl, Redundanzprinzipien, Selbsttests, Überwachung im Betrieb).

Die Vorgehensweise zur **Ermittlung des akzeptierten Risikos** wird in der CSM-Verordnung (EU) 402/2013 und (EU) 2015/1136 vorgeschrieben. Gängige Verfahren sind die Übernahme vorhandener Werte durch die Anwendung anerkannter Regeln der Technik oder der Vergleich mit einem bereits existierenden und akzeptierten Referenzsystem. Wenn beides nicht möglich ist, muss das akzeptierte Risiko explizit ermittelt werden (z.B. über die Verfahren Risikograph oder Risikomatrix).

Für technische Funktionen der Eisenbahnsicherungstechnik gilt die Vornorm DIN VDE V 0831-103 als anerkannte Regel der Technik. Dort werden die THR für typische Funktionen hergeleitet.

4) Festlegung der Systemanforderungen

Bei der **Festlegung der Systemanforderungen** werden die **funktionalen technischen Anforderungen an das System** vollständig **definiert**.

Dazu zählen beispielsweise:

- Funktionen
- Robustheit und Wartbarkeit
- Leistung und Effizienz
- Sicherheit
- Schnittstellen
- Betriebliche Anwendung
- Umgebungsbedingungen (vgl. oben; bspw. Klima, Mechanik physische und IT-Security)

Die Anforderungen müssen dabei eindeutig, vollständig, widerspruchsfrei korrekt, identifizierbar und prüfbar sein.

5) Architektur und Aufteilung der Systemanforderungen

In dieser Phase wird eine **Systemarchitektur** entwickelt, die in der Lage ist, die spezifizierten Anforderungen zu erfüllen. Die Anforderungen werden den Teilsystemen und

Komponenten zugewiesen und es werden die Schnittstellen zwischen diesen Teilsystemen und Komponenten spezifiziert.

6) Entwurf und Implementierung

In dieser Phase erfolgt der **Entwurf der Teilsysteme und Komponenten**, so dass sie die an sie gestellten Anforderungen erfüllen.

7) Herstellung

Es folgt die **Fertigung der Komponenten und Teilsysteme**.

8) Integration

In dieser Phase werden die **Teilsysteme und Komponenten zum Gesamtsystem integriert**. Es wird nachgewiesen, dass das integrierte System korrekt zusammenwirkt, um die vorgesehene Funktion zu erfüllen.

9) Systemvalidierung

Während der **Systemvalidierung** erfolgt die Prüfung und Bestätigung, dass das betrachtete System für den vorgesehenen Verwendungszweck geeignet ist und die festgelegten Anforderungen erfüllt. Die wesentlichen Mittel der Validierung sind Analysen und Tests.

10) Systemabnahme

Es folgt eine **abschließende Begutachtung bzw. Bewertung des Systems** und dessen Abnahme durch den Betreiber. Grundlage hierfür bilden die Nachweise der vorangegangenen Phasen und ggf. ergänzende Prüfungen oder Tests.

11) Betrieb, Instandhaltung und Leistungsüberwachung

Zu dieser Phase gehören:

- Betrieb des Systems
- Schulung des Personals
- Instandhaltung des Systems
- Fehlermanagement (Analyse, Bewertung, Korrektur)
- Führen des Gefährdungslogbuchs

12) Außerbetriebnahme

In der Phase Außerbetriebnahme erfolgen das **Abschalten des Systems** am Ende seiner Nutzung und das **Entfernen aus seiner Systemumgebung**.

(ii) Sicherheitsnachweis nach EN 50129

Im **Sicherheitsnachweis** wird **dargelegt und begründet, wie das System die gestellten Anforderungen an Funktionalität und Sicherheit erfüllt**. Aufbau und Struktur des Sicherheitsnachweises sind in der Norm EN 50129 festgelegt (siehe Abb. 6). Für Softwareanwendungen sind zusätzlich die Anforderungen der EN 50128 zu berücksichtigen.

Der „technische“ Nachweis erfolgt im Teil „Technischer Sicherheitsbericht“. Interessant wird sein, wie die wesentlichen normativ geforderten Inhalte an den Technischen Sicherheitsbericht auf eine KI-Anwendung angewendet werden können und welche

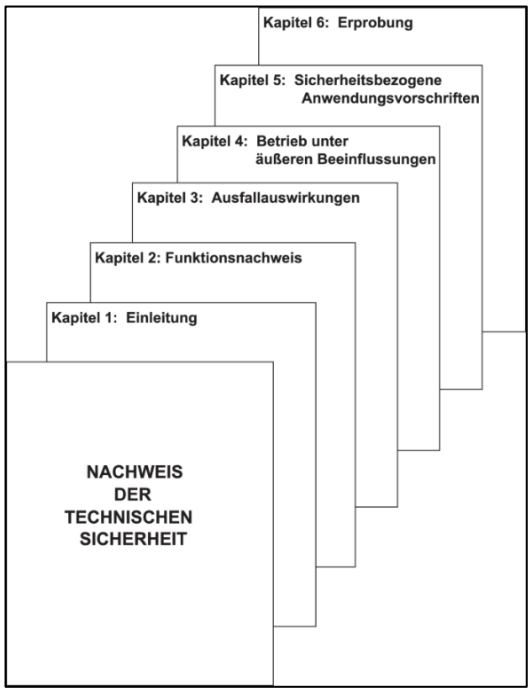


Abb. 3: Nachweis der funktionalen Sicherheit (DIN EN 50129)

Fragestellungen bzw. offenen Punkte aus Sicht der Nachweisführung gegenwärtig noch nicht hinreichend beantwortet sind (dazu 4 b ii).

1) Einleitung

In diesem Kapitel 1 erfolgt ein Überblick über das System und den Systementwurf und eine Darstellung der Prinzipien, auf denen sich die Sicherheit des Systems abstützt. Die Inhalte leiten sich dabei aus den Ergebnissen der Phasen „Systemkonzept“ bis „Festlegung der Systemanforderungen“ ab.

2) Nachweis des korrekten funktionalen Verhaltens

Der wesentliche Inhalt dieses Kapitels 2 ist der Nachweis zur Erfüllung der funktionalen Anteile der System-Anforderungsspezifikation (d.h. der funktionalen betrieblichen Anforderungen) und der Sicherheits-Anforderungsspezifikation (d.h. der funktionalen Sicherheitsanforderungen). Hierzu gehört der Nachweis der korrekten Hardware-Funktionalität und der Nachweis der korrekten Software-Funktionalität.

Der Nachweis der korrekten Softwarefunktionalität erfolgt durch Analysen und Tests.

3) Ausfallwirkungen

In Kapitel 3 ist vor allem die Ungefährlichkeit von Hardwareausfällen nachzuweisen.

4) Betrieb mit externen/äußeren Einflüssen

Kapitel 4 enthält den Nachweis, dass das entwickelte System auch unter Einwirkung der anzunehmenden externen Einflüsse seine betrieblichen Anforderungen und seine Sicherheitsanforderungen erfüllt. Externe bzw. äußere Einflüsse sind z.B. Umgebungsbedingungen wie Temperatur, Luftfeuchte, mechanische und elektrische Einflüsse. Je nach Art der verwendeten Sensoren kommen ggf. auch optische oder akustische Einflüsse hinzu.

5) Sicherheitsbezogenen Anwendungsbedingungen

Die sicherheitsbezogenen Anwendungsbedingungen sind in Kapitel 5 manifestiert. Hier erfolgt die Definition aller Regeln, Bedingungen und Einschränkungen, die bei der Anwendung des Systems zu beachten sind, damit dessen Sicherheit gewährleistet ist.

6) Ergebnisse der Sicherheitserprobung

Im Anschluss an den theoretischen Nachweis durch Analysen und Tests erfolgt die Sicherheitserprobung, dargestellt in Kapitel 6. Sie dient dazu, das Vertrauen in das neu entwickelte System zu stärken. Ihr Ziel ist eine hinreichende Erprobung des Systems unter allen relevanten Betriebsbedingungen. Zu beachten ist, dass die Sicherheitserprobung nicht mit den Tests zum Nachweis des korrekten funktionalen Verhaltens und

des Betriebs mit externen Einflüssen gleichzusetzen ist. Diese Tests müssen vor Beginn der Erprobung erfolgreich abgeschlossen sein. Man unterscheidet eine Sicherheitserprobung mit und ohne Sicherheitsverantwortung des zu erprobenden Systems.

(iii) Zusammenfassung

Nachdem nun abstrakt die Voraussetzungen zur Zertifizierung einzelner Teilsysteme dargestellt wurden, soll nachfolgend eine Fokussierung auf die im Projekt gestellte Forschungsfrage (bahnspezifische Anwendungen) erfolgen.

TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Projekt KI-bezogene Test- und Zulassungsmethoden

AP III: Verlässlichkeit und KI – Zentrale Begriffe

Abschlussbericht

Matthias Werner und Martin Boesler

20. Dezember 2021 (Version 0.95)

Inhaltsverzeichnis

1. Problem und Vorgehen

Dieser Report ist im Rahmen des Forschungsprojekts „KI-bezogene Test- und Zulassungsmethoden“, welches im ein Projekt des des *Smart Rail Connectivity Campus* und somit im Bahnbereich angesidelt ist, entstanden. Ziel des Gesamtprojektes ist es, Ansätze aufzuzeigen, wie Methoden der Künstlichen Intelligenz im sicherheitskritischen Bereich der Bahn zum Einsatz kommen können unter Berücksichtigung nicht nur der technischen, sondern auch der juristischen Gegebenheit, und ggf. Vorschläge zur Weiterentwicklung beider Bereiche zu machen.

Da es z. T. eine Verwirrung um die Begrifflichkeiten auf dem Gebiet der Künstlichen Intelligenz – insbesondere des maschinellen Lernens und Abgrenzung zu anderen Ansätzen der Künstlichen Intelligenz – und der Relevanz von algorithmischen Ansätzen (hier eben Ansätzen der Künstlichen Intelligenz) im Bezug auf die Systemverlässlichkeit und die funktionale Sicherheit gibt, war es die Aufgabe des Arbeitspakets III, Systeme der Künstlichen Intelligenz auf dem Blickwinkel der Verlässlichkeit zu betrachten, wesentliche Begriffe zu definieren und dabei die kritischen Punkte herauszuarbeiten. Da als ein mögliches Ergebnis des Gesamtprojektes eine Empfehlung zur Erarbeitung neuer Normen im Bereich der Zulassung von KI-Methoden betrachtet wird, wird in diesem Bericht nicht in erster Linie die Sicht des bestehenden Normenwerks (diese wird ohnehin in einem anderen Teilprojekt adressiert), sondern die Sicht des Standes der wissenschaftlichen Forschung eingenommen.

Dabei hat der Report folgende Struktur: Der erste Teil dieses Reports erörtert die Begriffe, Konzepte und ihre Beziehungen. Dabei wird im Abschnitt 2 Verlässlichkeit aus sich der Forschung diskutiert. Abschnitt 3 diskutiert das Konzept der Künstlichen Intelligenz, gibt eine grobe Taxonomie und erläutert grundsätzliche Ansätze für Lernverfahren in künstlichenen neuronalen Netzen. Auf der Grundlage der Abschnitte 2 und 3 werden im Abschnitt 4 die Besonderheiten von Systemen der Künstlichen Intelligenz im Bezug auf andere (sicherheitskritische) Systeme erarbeitet und auf die spezielle Rolle der Lernmusterdaten eingegangen. Dabei werden zwei Interpretationen für das Verständnis des Trainierens von Lernmusterdaten angeboten, die sich daraus ergebenden Konsequenzen diskutiert und auf die verschiedenen Auswirkungen auf die Verlässlichkeitsaspekte eingegangen. Der erste Teil schließt im einer Zusammenfassung der Ergebnisse in Thesenform.

Der zweite Teil dieses Reports ist ein Glossar der in diesem Kontext wesentlichen Begriffe.

Teil I.

Künstliche Intelligenz unter dem Gesichtspunkt der Verlässlichkeit

2. Verlässlichkeit

2.1. Konzept der Verlässlichkeit

In dem Projekt „KI-bezogene Test- und Zulassungsmethoden im Bahnbereich“ geht es um *sicherheitskritische* Systeme. Bei diesen Systemen machen der Test und die Zulassung einen wesentlichen Anteil an der gesamten Einführung aus. Sicherheitskritische Systeme werden in der Wissenschaft den *verlässlichen* Systemen zugerechnet. Daher wird hier zunächst kurz der Stand der Forschung im Gebiet Verlässlichkeit referiert.

Der Term „Verlässlichkeit“ (*dependability*) ist ein Konzept, das wiederum Mitte der 1980er Jahr aus dem damals sog. Gebiet der *Fehlertoleranten Systeme* entstand. Insbesondere die Arbeiten von LAPRIE und AVIŽIENIS haben hier Pionierarbeit geleistet, siehe z. B. [Lap85, AL86, Lap92, Lap95, LK96, ALR+01, ALR04, ALRL04].

> *Dependability* [Lap92] Dependability is defined as the **trustworthiness** of a computer system such that **reliance** can justifiable be placed **on the service it delivers.**
> The service delivered is its **behaviour as it is perceptible** to its user(s); a user is another system (human or physical) which interacts with the former.

Die Eigenschaft *Verlässlichkeit* wird dem *System* zugeschrieben. Im Folgenden wird – auch wenn es sich in der Regel um **eingebettete Systeme** handelt – stets unter System ein Computersystem oder der IT-Teil eines komplexeren Systems verstanden.

Da der Systembegriff zwar ein relativ abstraktes Konzept darstellt, aber der in der Verlässlichkeitsforschung genutzte Begriff etwas von den in anderen technischen Gebieten abweicht (und zwar bezüglich des *Zustands*), wird hier auch noch einmal der Systembegriff betrachtet.

2.2. Systembegriff

Ein *System* ist ein aus mehreren Einzelteilen zusammengesetztes Ganzes. Die Einzelteile sind Subsysteme und damit wiederum Systeme. Diese Rekursion endet bei *atomaren Systemen*. In der IT werden die Subsysteme meist *Komponenten* genannt und beispielsweise in Schichten (*Layern*) organisiert.

Jedes System besitzt einen *Zustand*. Alle von außen beobachtbaren Wirkungen des Systems gehören zum Zustand, man spricht vom *externen Zustand*. Über den externen Zustand interagiert das System mit seiner Umwelt. Die Subsysteme interagieren

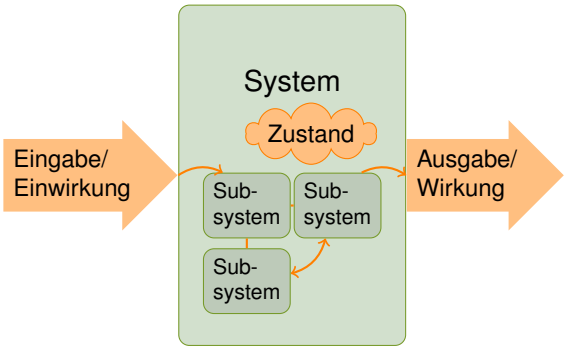


Abbildung 2.1.: System

miteinander – ebenfalls über ihren externen Zustand – und ggf. mit der Umwelt. Diese Wechselwirkung stellt den Dienst (*service*) des Systems dar, der dem Nutzern *users* des Systems zur Verfügung steht. Man beachte, dass sowohl der Dienstbegriff als auch der Nutzerbegriff hier sehr allgemein sind: Nutzer ist jeder Mensch oder jedes andere System, der/das mit dem System in Wechselwirkung steht. Dies schließt auch explizit Systeme oder Menschen ein, die nicht beabsichtigt mit dem System interagieren und damit unbeabsichtig vom Verhalten des Systems betroffen sind. Dabei werden alle Wirkungen des Systems als Dienst angesehen, egal ob nützlich oder schädlich, gewünscht oder unerwünscht. Dieser Unterschied wird erst bei den Attributen gemacht, die einem verlässlichen System zugewiesen werden:

Ein System hat eine *Struktur*, die bestimmt, wie seine Subsysteme interagieren können. Im Unterschied zum beispielsweise Automatenmodell wird die Struktur *als Teil des Zustands* betrachtet. Zu dem erweiterten Zustandsbegriff des verlässlichen Systems zählen nämlich auch Berechnungen, Kommunikation, gespeicherte Informationen, Verbindungen und physische Gegebenheiten. Anders als im „klassischen“ Zustandsstandsraum sind hier auch häufig bei anderen Systemkonzepten als „unveränderbar“ betrachtete Teile des Systems im Zustand enthalten. Jedoch werden bestimmte Änderungen des Zustand als unerwünscht angesehen. D. h., bei korrektem Verhalten ist der Zustandsraum eingeschränkt.

2.3. Attribute eines verlässlichen Systems

Mit dem im Abschnitt 2.2 gegebenen Systemkonzept können jetzt die ersten Attribute bzw. gewünschte Eigenschaften eines verlässlichen Systems definiert werden:

- Garantie der Bereitstellung des (nützlichen) Dienstes ⇒**Verfügbarkeit** (*availability*)
- Kontinuität des Dienstes ⇒**Zuverlässigkeit** (*reliability*)
- Verhinderung von Schaden ⇒**(technische) Sicherheit** (*safety*)

- Verhinderung von unzulässigen Zustandsänderungen ⇒**Integrität** (*integrity*)
- Möglichkeit zu Modifikationen und Reperaturen ⇒**Wartbarkeit** (*maintainability*)

Abbildung 2.2 gibt einen Überblick über diese Eigenschaften im Bezug auf das System.

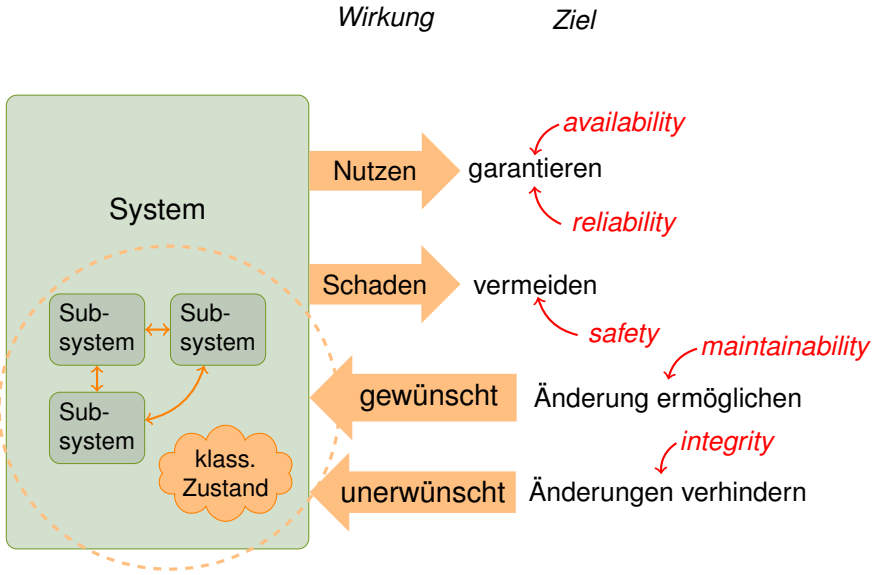


Abbildung 2.2.: Eigenschaften eines verlässlichen Systems (ohne Berücksichtigung von Befugnis)

Häufig muss unterschieden werden, wem ein (in der Regel: nützlicher) Dienst zur Verfügung gestellt wird. Man unterscheidet dabei befugte und nichtbefugte Nutzer. Letztere werden mitunter auch als *Nichtnutzer* bezeichnet. Bezüglich unterschiedlicher Dienste können unterschiedliche Nutzer befugt bzw. unbefugt sein. Da jedes Verhalten des Systems als Dienst angesehen wird, ist auch das unerwünschte Offenlegen von internen Zuständen ein (unerwünschter) Dienst.

Mit Bezug auf Befugnis bestimmter Nutzer zu bestimmten Aktionen können zwei der oben aufgeführten Eigenschaften als Eigenschaften der *Informationssicherheit* (*security*) aufgefasst werden. Hinzu kommt noch die Eigenschaft *Vertraulichkeit*:

- Verhindern von unbefugter Nutzung/Informationsoffenlegung ⇒**Vertraulichkeit** (*confidentiality*)

Die drei informationssicherheitsrelevanten Eigenschaften werden auch nach den Anfangsbuchstaben der englischen Bezeichnungen die CIA-Triade genannt.

Die genannten sechs Eigenschaften stellen alle Primäreigenschaften eines verlässlichen Systems dar. Abbildung 2.4 zeigt nochmals die Beziehung zwischen Verlässlichkeitseigenschaften im Allgemeinen und Eigenschaften der Informationssicherheit. LA-

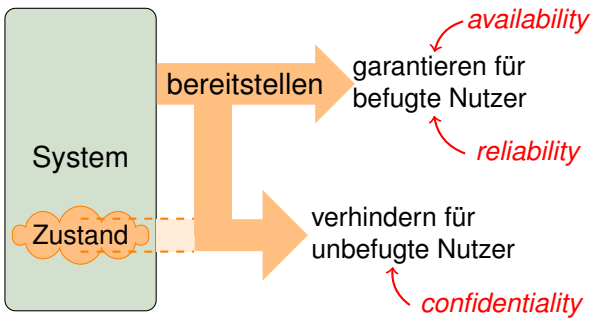


Abbildung 2.3.: CIA-Triade

PRIE und Co. sehen damit Informationssicherheit als Teilgebiet der verlässlichen Systeme insgesamt.

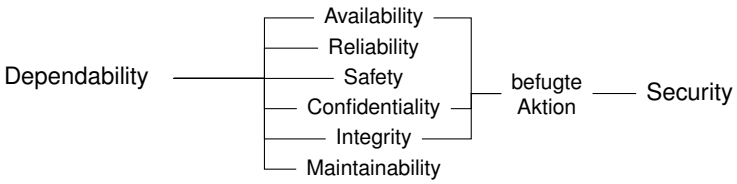


Abbildung 2.4.: Verlässlichkeit vs. Informationssicherheit

Neben diesen Primäreigenschaften gibt es noch sekundäre Attribute. Dazu zählen beispielsweise die Robustheit (*robustness*) oder die Zurechenbarkeit (*accountability*). Beides (und alle anderen Sekundärattribute) sind jedoch stets nur hilfsweise zu verwirklichen, um ein oder mehrere der der Primärattribute durchzusetzen.

Je nach Anwendungsfall kann es komplexe Wechselwirkungen zwischen den verschiedenen Eigenschaften geben. Es ist beispielsweise in der Regel einfach, die technische Sicherheit eines Systems zuungunsten der Verfügbarkeit zu erhöhen: Das System macht einfach nichts (und bewirkt damit auch keinen Schaden). Gleichfalls kann beispielsweise eine Beeinträchtigung der Integrität sowohl die Zuverlässigkeit, die Verfügbarkeit und die Vertraulichkeit stören. Umgekeht kann ein Bruch der Vertraulichkeit die Integrität beeinträchtigen, usw.

2.4. Beeinträchtigungen

Es kann unerwünschte (aber im Prinzip nicht unerwartete) Umstände geben, die die Verlässlichkeit beeinträchtigen, d.h. die Verlässlichkeitsziele bedrohen. Dann weicht der Dienst des Systems von der Erwartung ab. Der Dienst eines Systems ist *korrekt*, wenn er die Erwartungen erfüllt. Diese Erwartungen sind häufig (aber nicht immer und vor allem nicht immer vollständig) in der Spezifikation beschrieben. Allerdings können bereits Spezifikationen inkorrekt sein: Spezifikationen, die ein für Nutzer unakzeptables

Verhalten vorgeben (zum Beispiel weil das beschriebene Verhalten den guten Sitten oder gesetzlichen Vorgaben widerspricht), sind ggf. fehlerhaft.

Bei den Beeinträchtigungen unterscheidet man:

- Das **Ereignis**, dass ein Dienst beginnt von der Erwartung abzuweichen, heißt **(Dienst-)Versagen** (failure) oder **Ausfall**.
- Die Ursache für ein Versagen ist immer (mindestens) ein *Fehlerzustand* (*error*), d.h. Teil(e) des Gesamtzustands, die bei korrektem Verhalten nicht auftreten. Man beachte, dass es sich beim Error, im Gegensatz zum Versagen, um einen **Zustand** handelt. Hier ist ausdrücklich der erweiterte Zustandsbegriff entsprechend Abschnitt 2.2 gemeint.

  Fehlerzustände können ggf. wieder verschwinden, ohne ein Versagen zu bewirken. Sie können auch unentdeckt verweilen; dann spricht man von einem *latenten* Fehlerzustand.
- Ein Fehlerzustand hat (mindestens) eine **Fehlerursache** (*fault*). Jede derartige Ursache liegt im Versagen eines anderen Dienstes (wobei der Mensch auch als Diensterbringer aufgefasst wird: Systementwicklung kann als Dienst betrachtet werden.) Um diese Rekursion zu brechen, definiert man „Fehlerursache“ als tatsächliche *oder angenommene* Ursache eines Fehlerzustands.

Man beachte, dass die Korrektheit des *Dienstes* und nicht des *Systems* definiert wird, da es in dieser Sichtweise[1] kaum fehlerfreie Systeme gibt; nur Systeme, die *noch nicht* versagt haben.

Letztendlich kann als letzte Fehlerursache stets der Mensch angesehen werden, aufgrund des Unvermögens oder der Ablehnung alle Situationen vorherzusehen. Diese Sichtweise ist jedoch in der Praxis nicht immer hilfreich um verlässliche Systeme zu schaffen. Daher gibt es umfangreiche Taxonomien von Fehlerursachen und -wirkungen, die in erster Linie der Kommunikation in Entwurfsprozess dieser Systeme dienen. Als grobe Kategorisierung dient oft die Einteilung in die folgenden drei (teilweise überlappende) Gruppen:

- **Entwicklungsfehler** werden während Systementwicklung, -veränderung oder der Geschäftsprozessentwicklung und -einführung erzeugt
- **Physische Fehler** sind alle Fehlerursachen, die die Hardware betreffen
- **Interaktionsfehler** sind alle externen Fehlerursachen

2.5. Metriken

Die verschiedenen im Abschnitt 2.3 diskutierten Eigenschaften sind zunächst *qualitativ* aufzufassen. Sie werden jedoch quantitativ durch verschiedene Metriken untersetzt.

[1] von LAPRIE und anderen, siehe z.B. [LK96]

Die Unterscheidung zwischen Eigenschaft und Metrik wird z. T. in der eher technischen Literatur verwischt, indem unmittelbar auf die Metrik Bezug genommen wird, obwohl allgemein die Eigenschaft behandelt wird. Zu der Verwirrung trägt bei, dass einige Metriken die gleiche Bezeichnung haben wie die Eigenschaften, die sie quantifizieren. Beispielsweise wird die Wahrscheinlichkeit, dass ein System über einen Zeitraum $[t_0, t_0 + t]$ ein korrektes Verhalten zeigt, wenn es dies bereits zum Zeitpunkt t_0 tat, ebenfalls Zuverlässigkeit (*reliability*) genannt.[2] Es gibt aber auch noch andere Metriken, die Eigenschaft Zuverlässigkeit (also die Kontinuität des Dienstes) quantitativ beschreiben, beispielsweise die *mittlere Zeit bis zum Ausfall* (*mean time to failure*) $MTTF$.

Solche Verwechslungen von Metrik und Eigenschaft finden sich sogar in für dieses Projekt relevanten Normen: So findet man in [ICE00] für *Instandhaltbarkeit* die Definition:

> Wahrscheinlichkeit dafür, dass für eine Komponente unter gegebenen Einsatzbedingungen eine bestimmte Instandhaltungsmaßnahme innerhalb einer festgelegten Zeitspanne ausgeführt werden kann, wenn die Instandhaltung unter festgelegten Bedingungen erfolgt und festgelegte Verfahren und Hilfsmittel eingesetzt werden.

Eine Wahrscheinlichkeit ist aber in jedem Fall bereits eine Metrik. Die Vermischung von Eigenschaft und Metrik ist kritisch, da einerseits verschiedene Metriken verschiedene Aspekte der Eigenschaft zum Ausdruck bringen können und da andererseits nichtanwendbare Metriken zu Missverständnissen über die Eigenschaft führen können.

2.6. Nachweis von Verlässlichkeitseigenschaften

Um den Nachweis von Verlässlichkeitseigenschaften eines Systems zu führen, muss

1. die Korrektheit des Verhaltens des Systems unter der Voraussetzung der Erfüllung spezifischer Annahmen nachgewiesen werden und
2. Betrachtungen zur Belastbarkeit dieser Annahmen durchgeführt werden.

Bei der Belastbarkeit geht es um die Dinge wie die Abdeckung von Fehlermodellen, probabilistische Lastannahmen oder Ausfahlraten von Hardware. Obwohl diese Betrachtungen bei KI-basierten Systemen sehr komplex sein können, unterscheiden sie sich nicht prinzipiell von denen bei anderen IT-Systeme. Anders verhält es sich beim Punkt 1, dem Nachweis der Korrektheit des Verhaltens. Dieser Nachweis wird *Verifikation* (*verification*) genannt.

Prinzipiell gibt es drei grundsätzliche Möglichkeiten zur Verifikation von softwarebasierten Systemen:

1. **Testen**: Das System erhält ausgewählte Eingabedaten. Es wird überprüft, ob die Reaktion des Systems dem in der Spezifikation des System angegebenen Verhalten entspricht.

[2] Eine andere Bezeichnung für diese Metrik ist *Überlebenswahrscheinlichkeit*.

2. **Auditieren**: Die Implementationen des Systems werden auf das Vorhandensein typischer Schwachstellen untersucht.

3. **Formale Verifikation**: Mit Hilfe von mathematischen Methoden wird die Übereinstimmung des in der Spezifikation angegeben Verhaltens mit dem Verhalten des Systems (/der Systememplementation) bewiesen.

Man beachte, dass durch Testen und Auditieren genau genommen *kein* vollständiger Nachweis einer Korrektheit durchgeführt werden kann. Es kann zwar das Vorhandensein von Fehlern nachgewiesen werden, jedoch nicht ihre Abwesenheit.[3]

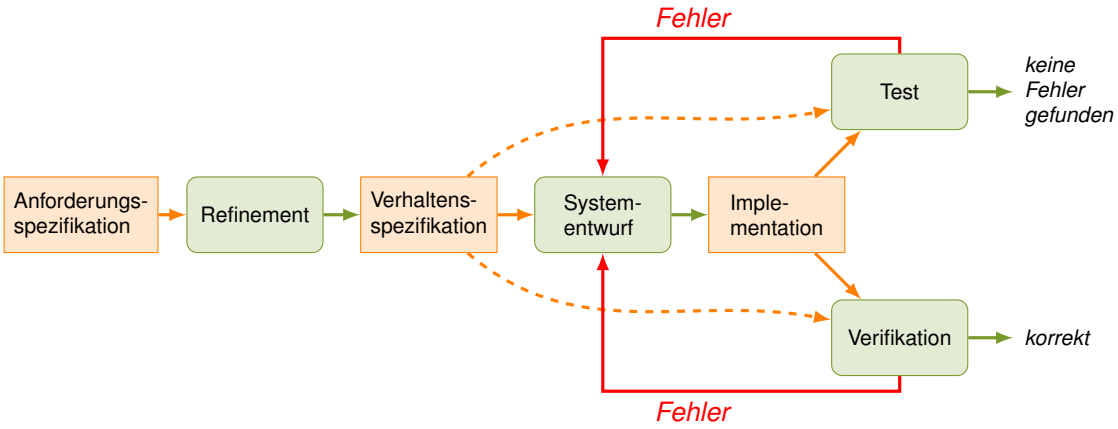


Abbildung 2.5.: Unterschiede zwischen Testen und formaler Verifikation

[3] Eine Ausnahme bildet das *erschöpfende Testen*, bei dem alle nur möglichen Eingaben in den Test einbezogen werden. Dies ist aber nur bei einer endlichen und dazu nicht zu großen Anzahl von Eingabevarianten möglich, was in der Praxis sehr selten vorkommt.

3. Künstliche Intelligenz (KI)

3.1. Begriff der Künstlichen Intelligenz

In diesem Abschnitt sollen Grundkonzepte der Künstlichen Intelligenz erläutert werden. Dabei ist die Abgrenzung der Künstlichen Intelligenz zu anderen Gebieten außerordentlich schwierig. Zur Demonstration seien hier drei Definitionsansätze dokumentiert, die auch bis zu einem gewissen Grad die Entwicklungsgeschichte der Künstlichen Intelligenz widerspiegeln:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.
>
> J. MCCARTHY, 1955 nach [MRSM06]

> The study of how to make computers do things at which, at the moment, people are better.
>
> RICH, 1983 [Ric83]

> The art of creating machines that perform functions that require intelligence when performed by people.
>
> KURZWEIL, 1990 [Kur90]

> Artificial intelligence, broadly (and somewhat circularly) defined, is concerned with intelligent behavior in artifacts. Intelligent behavior, in turn, involves perception, reasoning, learning, communicating, and acting in complex environments.
>
> NILSSON, 1998 [Nil98]

Man merkt diesen Definitionen an, dass Künstliche Intelligenz ein Gebiet ist, das sich wohl nur sehr schwer abgrenzen lässt oder sich sogar einer Definition entzieht, auch weil bereits das Konzept der *Intelligenz* nicht vollständig geklärt ist. Entsprechend ist den Autoren keine wissenschaftlich allgemein anerkannte, diskriminierende Definition von Künstliche Intelligenz bekannt. Für dieses Projekt wurde daher eine Arbeitsdefinition entwickelt, die keinen Anspruch auf Vollständigkeit oder hinreichende Abgrenzung erhebt, jedoch für dieses Projekt wesentliche Aspekte einschließt:

> *Künstliche Intelligenz* Künstliche Intelligenz ist die Einheit von Hard- und Softwarekomponenten, welche – inspiriert von biologischen Vorbildern – zum Zweck der autonomen, effizienten und kreativen Problemlösung konzipiert, konstruiert und eingesetzt wird.

Häufig unterscheidet man in der Künstliche Intelligenz zwischen *starker* und *schwacher Künstliche Intelligenz*. Mit starker KI wird etwas bezeichnet, das die menschliche Intelligenz nachbildet. Dabei gilt es als offene Forschungsfrage, ob dies auch ein Bewusstsein notwendig macht oder sich dieses als Folge ergibt. Schon die Frage wie entschieden werden kann, ob etwas die menschliche Intelligenz nachbildet, ist nicht geklärt. Der erste Vorschlag dazu stammt bereits von Alan Turin aus dem Jahr 1950 (Turing-Test, [Tur50]), ist aber umstritten. Eine Reihe anderer Tests wurde vorgeschlagen, so z. B. der Lovelance-Test [BBF03] oder die Winograd-Challenge [J.L14]. Nach allgemeiner Ansicht ist bisher noch keine starke Künstliche Intelligenz erreicht; mitunter wird sogar postuliert, dass sie unerreichbar ist. Trotzdem ist die öffentliche Vorstellung von KI vielfach durch die künstlerische Aufarbeitung (Film, Literatur) des Konzepts der „starken KI“ geprägt.

Die *schwache Künstliche Intelligenz* beschäftigt sich dagegen mit konkreten einzelnen Anwendungsproblemen des menschlichen Denkens bzw. der menschlichen Kognition. Alle heutigen KI-Anwendungen, und erst Recht die in diesem Projekt relevanten Anwendungen, liegen in der schwachen Künstliche Intelligenz.

3.2. Methoden der Künstlichen Intelligenz

Im Laufe der Geschichte der Künstlichen Intelligenz wurden eine fast unüberschaubare Vielzahl von Ansätzen und Methoden entwickelt. Für dieses Projekt haben wir versucht, eine Taxonomie der Ansätze aufzustellen und dabei drei Hauptrichtungen identifiziert:

- Symbolische/regelbasierte Verfahren
- Klassische/transparente lernende Verfahren
- Neuroinspirierte Verfahren

Die neuroinspirerten Methoden können noch einmal in drei Gruppen unterteilt werden: unüberwachtes, überwachtes und bestärkendes Lernen.

3.2.1. Regelbasierte Verfahren

Bei den *regelbasierte* Verfahren wird Wissen explizit vorgegeben. Dieses wird in Regeln mit einer geeigneten formalen Sprache ausgedrückt. Mögliche Unsicherheiten bzw. Unschärfen müssen hier explizit gemacht werden. Die gegebenen Regeln werden durch

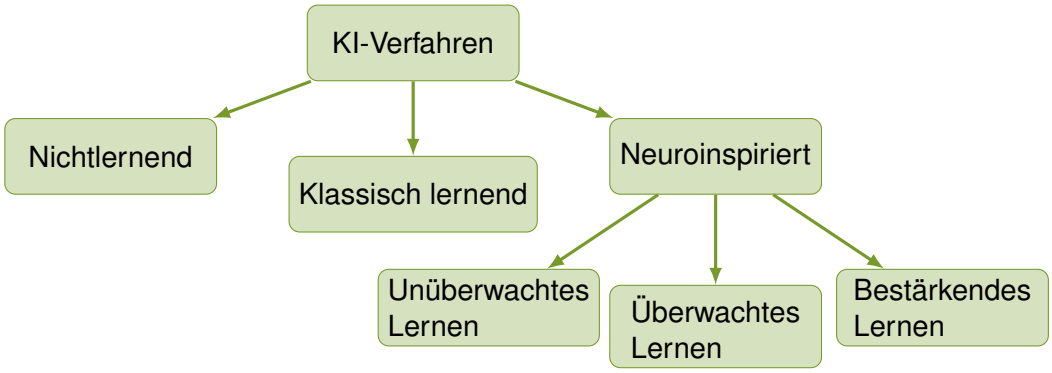


Abbildung 3.1.: Klassifikation von KI-Ansätzen

die Maschine auf syntaktischer Ebene erfasst und auf die gegebene Datenbasis angewendet. Man spricht hier auch häufig vom fallbasierten Schließen (*case-based reasoning, CDR*). Diese Anwendungen sind – jenseits einer möglichen Komplexität – für den Menschen transparent (bzw. können leicht transparent gemacht werden) und sind, solange die Software korrekt implementiert ist und die Hardware keine Ausfälle zu verzeichnen hat, fehlerfrei innerhalb der vorgegeben Datenbasis und Regeln. Natürlich ist es möglich, dass die Regeln oder die Datenbasis nicht das Beabsichtigte ausdrücken; jedoch sind sowohl Datenbasis als auch Regeln leicht überprüfbar. Man nutzt daher solche Verfahren zur Überprüfung oder zur Erstellung von formalen Beweisen. Die eigentlichen Algorithmen in diesen Verfahren sind in der Regel clevere Suchalgorithmen.

Beispiele für diese Klasse der KI-Verfahren sind die Programmiersprache Prolog oder Programme zum Umgang mit symbolischer Mathematik, wie z. B. Mathcad.

Im Gegensatz zu den in den Abschnitten 3.2.2 und 3.2.3 genannten Verfahren handelt es sich nicht um lernende Verfahren.

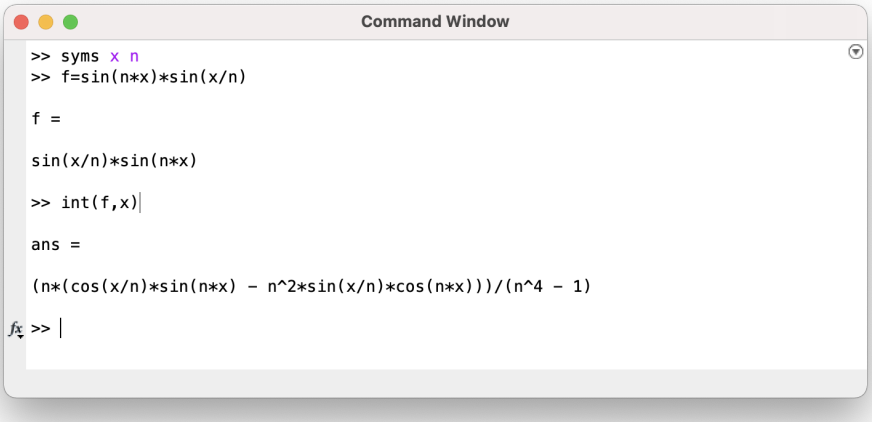
3.2.2. Klassische lernende Verfahren

Lernende Verfahren – man spricht hier von *maschinellem Lernen* (ML) –, egal ob klassisch lernende oder neuroinspirerte Verfahren (siehe Abschnitt 3.2.3), beruhen auf Prinzipien der mathematischen Optimierung.

Bei den klassischen Verfahren gibt es in der Regel ein in den Strukturen vorgegebenes Modell, dessen Parameter instanziiert bzw. optimiert werden. Dazu werden Daten und Informationen aus der Lerndomäne vorgegeben. Mit Hilfe einer Zielfunktion werden die Modellparameter angepasst, das akkumulierte Wissen im Modell konserviert. Klassisch soll hier nicht implizieren, dass es sich in jedem Fall um ältere Verfahren als bei den im nächsten Abschnitt (3.2.3) diskutierten handelt. Im Gegenteil, bereits in den 1940er Jahren (vgl. [Kri07]) gab es Ansätze für künstliche neuronale Netzwerke. Der eigentliche Gegensatz zu den neuroinspirierten Verfahren ist, dass eine gemachte Modellvorhersage bei den klassischen lernenden Verfahren unmittelbar nachvollziehbar

```
% Wenn X Vater von Z ist und Z  Vater von Y ist, dann ist X Großvater von Y
grossvater(X, Y) :-
vater(X, Z),
vater(Z, Y).
% Adam ist der Vater von Tobias
vater(adam, tobias).
% Tobias ist der Vater von Frank
vater(tobias, frank).
% Abfrage, ob Adam der Großvater von Frank ist
?- grossvater(adam, frank).
true.
```

(a) Prolog



(b) Mathcad

Abbildung 3.2.: Beispiele für regelbasierte KI-Anwendungen

bzw. nachprüfbar ist.

Beispiele für klassische lernende Verfahren sind lernende Regelungen in der Steuerungstechnik oder Anwendungen des Entscheidungsbaum-Lernens (*decicion tree learning*).

3.2.3. Neuroinspirierte Verfahren

Die neuroispirierten Verfahren nehmen sich Systeme der biologischen Informationsverarbeitung zum Vorbild: neuronale Netze. In sogenannten künstlichen neuronalen Netzen[1] (KNN) werden Elemente der biologischen Vorlagen abstrakt nachgebildet. Kernelement ist dabei ein künstliches Neuron als Element der Informations- bzw. Signalverarbeitung.

[1] Im Folgenden ist mit „neuronales Netz“ stets ein künstliches neuronales Netz gemeint. Gleiches gilt für „Neuron“.

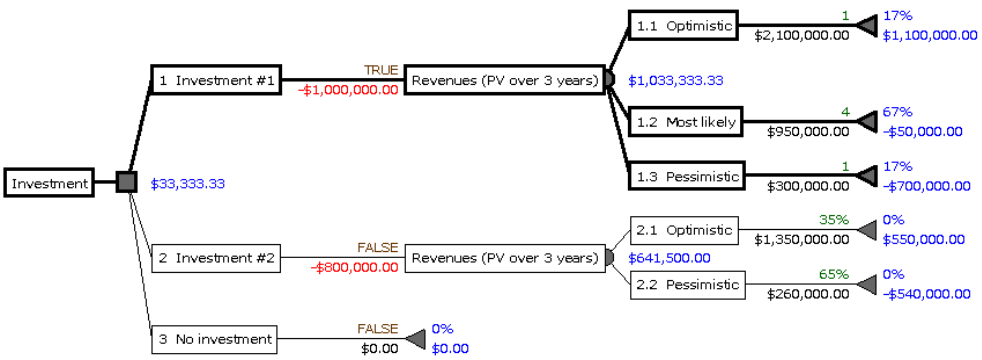


Abbildung 3.3.: Beispiel für klassische lernende Verfahren: Decision Tree Learning (Bildquelle: Wikimedia)

Im Prinzip kann ein künstliches Neuron mit einer Vektorfunktion $\vec{y} = f(\vec{x})$ beschrieben werden, wobei $\vec{x}$ das (mehrdimensionale) Eingangs- und $\vec{y}$ das (ebenfalls mehrdimensionale) Ausgangssignal darstellt. Jedoch haben die Neuronen einen *Aktivierungszustand*, der sich während des Betriebs ändert. Man kann also von parametrischen Funktionen reden.

Im Detail sind jedoch mehrere Funktionen zu unterscheiden, die jeweils einen Beitrag zur Wirkungsweise eines Neurons liefern:

- *Propagierungsfunktion*: Die Propagierungsfunktion verwandelt den Eingangsvektor in einen Skalarwert. Dabei entscheidet ein *Schwellenwert*, ob das Neuron überhaupt aktiviert wird.
- *Aktivierungsfunktion*: Die Aktivierungsfunktion bestimmt aus dem Eingangsskalar und dem alten Aktivierungszustand einen neuen Aktivierungszustand.
- *Ausgabefunktion*: Die Ausgabefunktion berechnet aus dem Aktivierungszustand die Werte, die als Ausgabewertevektor zurückgegeben werden.

Die Neuronen sind gekoppelt und bilden eben das neuronale Netz. Es gibt verschiedene typische Netztopologien, die jeweils spezifische Eigenschaften besitzen. Einige der Neuronen werden mit Eingangssignalen verbunden. Ausgänge der Neuronen werden mit den Eingängen anderer Neuronen verbunden oder stellen das Ausgangssignal des Netzes dar. In diesem Zusammenhang bezeichnet der Begriff *deep learning* neuronale Netze, die mehr als eine Schicht von Neuronen zwischen Eingabe und Ausgabe aufweisen.

Beim Lernen (oder *Trainieren*) werden dem Netz Signalmustern eingegeben. Während des Lernens wird dabei das neuronale Netz angepasst. Dafür gibt es unterschiedliche Möglichkeiten. Es können

- neue Verbindungen entwickelt,

- vorhandene Verbindungen gelöscht,
- Verbindungsgewichte verändert,
- Schwellenwerte von Neuronen geändert,
- Aktivierungs-, Propagierungs- und/oder Ausgabefunktion abwandelt,
- neue Neurone entwickelt/vorhandene Neuronen gelöscht

werden.

Mit Hilfe von neuronalen Netzen können verschiedene Ansätze des Lernens – man spricht vom Trainieren des Netzes – realisiert werden:

Unüberwachtes Lernen. Beim unüberwachtem Lernen werden dem Netz nur Eingabemuster vorgegeben, aber keine Lernhilfen oder Bewertungsvorgaben. Das Netz kann selbständig eine Einteilung der Eingabemuster vornehmen. Dies entspricht einer Abstraktionsleistung in biologischen Systemen. Typische Anwendungsfälle für diese Art des Lernens sind Fragestellungen der Zusammengehörigkeit, z.B. Clusterbildungen. Unüberwachtes Lernen wird häufig zur Dimensionsreduktion von Eingabedaten benutzt, die dann anderweitig weiterverarbeitet werden.

Abbildung 3.4.: Beispiel für Clustering (aus [Kri07])

Überwachtes Lernen. Beim Überwachtem Lernen (*supervised learning*) werden dem Netz zu den Eingabemustern die gewünschten Ergebnisse vorgegeben. Damit kann das Assoziieren des biologischen Vorbilds nachempfunden werden. Dies ist zielgerichteter als das unüberwachte Lernen, setzt aber das Vorhandensein der korrekten Abstraktion bereits voraus.

Bestärkendes Lernen. Bestärkendes Lernen kann als ein Mittelweg zwischen unüberwachtem überwachtem Lernen angesehen werden. Wieder werden Lernmuster vorgegeben. Es werden zwar keine Lernziele gegeben, aber die Reaktion des Netzes wird bewertet. Dies erlaubt es auch gute Lösungen zu finden, die außerhalb des erwarteten Bereichs waren. Auch wenn nach allgemeiner Meinung der

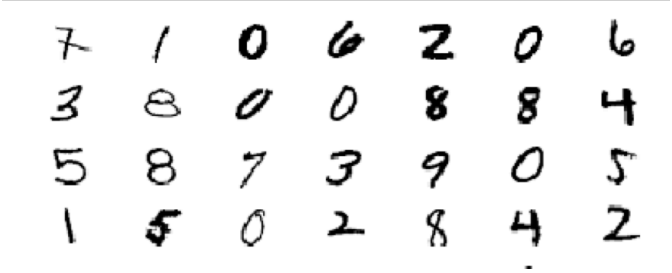


Abbildung 3.5.: Beispiel für überwachtes Lernen: Schrifterkennung

Kognitionsissenschaftler bei der menschlichen Kreativität andere Prozesse ablaufen, könnte man im übertragenen Sinn davon sprechen, dass beim bestärkenden Lernen Raum für kreative Losungen geschaffen wird.

Typische Anwendungsfälle sind koplexe Entscheidungssituationen wie z. B. in strategischen Spielen wie Schach oder Go.

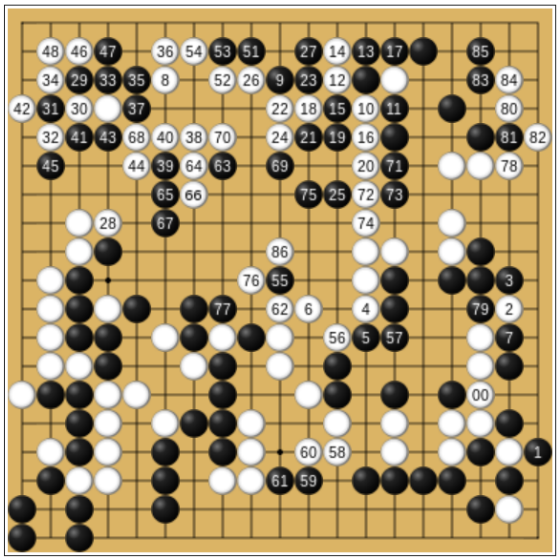


Abbildung 3.6.: Beispiel für bestärkendes Lernen: AlphaGo-Partie gegen Tang Weixing (31. Dezember 2016), die vom Computer gewonnen wurde (Bildquelle: Wikimedia)

3.3. Einsatz von Künstlicher Intelligenz

Auch wenn alle im voherigen Abschnitt diskutierten Ansätze zur künstlichen Intelligenz gezählt werden, wird heute der Begriff Künstliche Intelligenz z. T. synonym zu maschi-

nellen Lernen bzw. sogar meist zu Einsatz von KNN angesehen.[2] Auch die Motivation für das laufende Projekt waren in erster Linie künstliche neuronale Netze.

Wenn es zur Bearbeitung von Problemen herkömmliche – also nicht-KNN-basierte – Lösungsansätze gibt, schneiden diese derzeit im direkten Vergleich fast immer qualitativ besser ab als künstliche neuronale Netze. Daher gibt es für künstliche neuronale Netzeeigentlich nur zwei sinnvolle Einsatzgebiete:

1. Für Probleme, für die keine *effiziente* Lösungsstrategie bekannt sind, z. B. wegen der Komplexität des Problems und der damit verbundenen Rechenzeit und/oder dem Ressourcenverbrauch. Typische Vertreter dieser Probleme sind strategische Spiele wie Schach oder Go.

2. Probleme, bei denen die Problembeschreibung nicht oder nicht vollständig auf einer hinreichend formalen Ebene erfolgen kann, z. B. wegen fehlender oder unscharfer Informationen oder der Unschärfen bei der Benutzung abstrakter Konzepte. Typische Beispiele sind hier Objekterkennung für generalisierte Kategorien von Objekten.

Beide Einsatzbereiche können sich überschneiden, sind aber im Wesen unterschiedlich und verlangen im Einzelnen einen unterschiedlichen Umgang. Dies gilt insbesondere für alle Aspekte im Entwicklungs- und Lebenszyklus einer Anwendung, bei der es auf die Problemspezifikation ankommt. Im ersten Einsatzgebiet ist die Spezifikation häufig in einer formalen Exaktheit angebbar, dass das Verhalten des Systems unmittelbar (und prinzipiell vollständig) gegenüber der Spezifikation geprüft werden kann. Im zweiten Einsatzbereich ist dies nicht der Fall.

[2] Während der sogenannten ersten Blüte der KI von Mitte der 1950er Jahre an dominierten dagegen die symbolischen Verfahren.

4. Diskussion: Besonderheiten von KI-basierten Systemen im Bezug auf Verlässlichkeit

4.1. Abgrenzung zu anderen Systemdomänen

In diesem Abschnitt soll auf der Grundlage der in den Abschnitten und 2 und 3 vorgestellten Konzepte und Begriffe diskutiert werden, worin sich KI-basierte Systeme von anderen, ähnlichen aber nicht-KI-basierten, Systemen im Bezug auf die Verlässlichkeitsaspekte unterscheiden. Während im Abschnitt 3 im Wesentlichen auf den Aspekt der Funktionsweise und der Anwendung eingegangen wurde, wird hier eine technische Einordnung benötigt. Vom technischen Standpunkt aus sind Systeme der Künstlichen Intelligenz komplexe, aus IT-Hardware und Software bestehende Systeme, die in einem sicherheitskritischem Kontext eingesetzt werden.

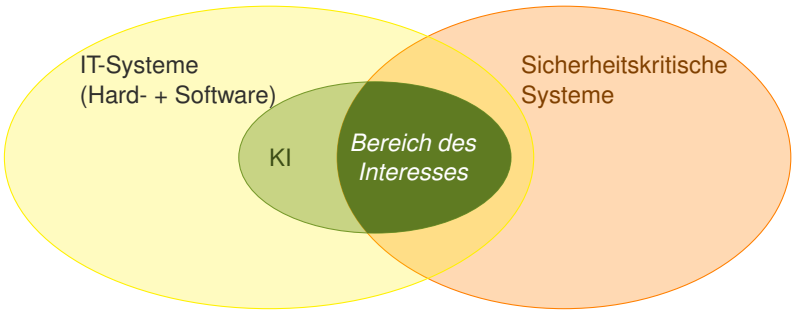


Abbildung 4.1.: Problemabgrenzung

Mit anderen Worten: Die Probleme, die bei anderen IT-Systemen im kritischen Einsatz existieren, existieren für die hier interessanten Systeme auch, zeichnen sie aber nicht aus. Im Kontext der Zulassung von solchen Systemen wird das Konzept der *funktionalen Sicherheit* (*functional safety*) genutzt, was den Begriff der technischen Sicherheit (siehe Abschnitt 2.3) untersetzt:

> *Functional Safety* [IEC02, ISO18] Functional safety is the absence of **unreasonable risk** due to hazards caused by malfunctioning **behavior** of systems. [*Hervorhebungen der Autoren*]

Man beachte, dass hier zwei Aspekte angesprochen sind: das Risiko, was sich im Kontext der Nutzer bzw. der Umgebung des Systems realisiert, und das Verhalten des

Systems selbst. Ersteres ist jedoch unabhängig von der Funktionsweise des Systems, sondern ist durch den EInsatz des Systems geprägt. Bezüglich des Risikos ist das System eine Blackbox, ein Implementierungsmethode wie KI spielt hier keine Rolle. Anders ist es bei beim Systemverhalten und dessen Korrektheit: dies ist explizit durch die Künstliche Intelligenz geprägt.

Allerdings gibt es bei den nichtlernenden und (größtenteils) bei den klassisch lernenden Verfahren keinen prinzipiellen Unterschied zu anderen (komplexen) IT-Systemen. Offensichtlich ist dies bei den regelbasierten KI-Systemen: sie unterscheiden sich weder im Prozess der Softwareerstellung[1] noch in der Nutzung grundsätzlich von Systemen, die nicht KI-basiert sind. Entsprechend greifen hier alle bereits bekannten Ansätze des Testens und der Zulassung.

Bei Systemen des maschinellen Lernens unterscheidet sich der Prozess der Softwareentwicklung: Er ist erst nach der Lernphase abgeschlossen. Während aber bei den meisten klassischen Verfahren des machinellen Lernens die Struktur des Modells vorgegeben ist und das Lernen damit eher einer Art Parameterisierung entspricht, stellt sich bei den neuroinspirierten Verfahren die Frage, wie der Trainingsprozess aus der Sicht der Softwareentwicklung und insbesondere der Verifizierung eingeordnet werden kann.

4.2. Die Rolle der Lernmusterdaten im Softwareentwicklungsprozess

Da ein neurales Netz erst nach dem Training (siehe Abschnitt 3.2.3) produktiv eingesetzt werden kann, wird das Trainieren ein Teil des Entwicklungsprozesses. Um einen Vergleich mit bereits bekannten Methoden zur Verifikation zu ermöglichen, stellt sich die Frage, wie das Trainieren bzw. die Lernmusterdaten zu interpretieren sind. Dabei bieten sich zur Interpretationen zwei Analogien an:

- **Analogie modellgetriebene Entwicklung: Daten sind High-Level-Code.** In dieser Interpretation werden die Lernmusterdaten als Instanziierung eines Modells (Modellparameter) aufgefasst. Dies setzt eine Analogie zum Ansatz der *modellgetriebenen Entwicklung* (*model driven development*, MDD). In der modellgetriebenen Entwicklung werden Modellsprachen eingesetzt, deren Abstraktionen eine Nähe zur Anwendungsdomäne haben. Die Untersetzung auf maschinennahe Sprachen erfolgt automatisiert.

  In dieser Interpretation entsprächen die Lernmusterdaten dem High-Level-Code der modellgetriebenen Entwicklung. Jedoch ist eine der Grundlagen der modellgetriebenen Entwicklung eine exakte Semantik der Modellsprache. Damit kann die Verifikation zweistufig verlaufen:

  1. Nachweis, dass die Übersetzung der Modellsprache in die maschinennahe Darstellung korrekt ist;

[1] Bezüglich der Hardwareentwicklung gibt es ohnehin keine Unterschiede.

2. Nachweis der Korrektheit des Modells auf der Ebene der Modellsprache.

Für klassische Ansätze des maschinellen Lernens ist ein solches Vorgehen durchaus möglich. Beispielsweise werden durch das Lernen bei lernenden Regelungsalgorithmen lediglich Unschärfen eliminiert, deren semantische Bedeutung jedoch klar ist.

Wenn jedoch keine klare Semantik der Modellsprache bekannt ist, wie dies bei den neuroinspirierten Verfahren der Fall ist, müsste die semantischen Information der Lernmusterdaten überprüft werden, z. B. durch ein entsprechendes Modell. Wenn es jedoch gelänge, ein solches Modell aufzustellen, wäre äquivalent zum dem durch das Lernen geschaffene Modell der Einsatz eines neuronalen Netzes überflüssig.

- **Analogie Programmierung: Trainieren ist Programmieren.** In dieser Interpretation wird das Trainieren mit dem Vorgang der Programmierung eines Systems durch einen Menschen gleichgesetzt. Damit entzieht sich der Übersetzungsprozess der Verifikation. Vielmehr muss das Ergebnis verifiziert werden. In diesem Fall haben die Lernmusterdaten die Funktion einer *Spezifikation* des Verhaltens, und das Trainieren ist ein ein Vorgang der Codeerstellung.

  Ein Verhaltensspezifikation untersetzt (gegebenenfalls über mehrere Refinementschritte) eine *Anforderungsspezifikation*. Das erste, allerding relativ geringe Problem dieses Ansatzes ist, dass Daten, die als Spezifikation aufgefasst werden, nicht mehr zur Verifikation der Anwendungsfunktionalität des neuronalen Netzes herangezogen werden können. Man behilft sich hier dadurch, dass nur eine Teilmenge der Lernmusterdaten auch tatsächlich zum Trainieren herangezogen wird, während der Rest zur Verifikation genutzt wird. Dies funktioniert jedoch nur unter der Annahme, dass beide Teilmengen die gleichen Informationen tragen.

  Das größere Problem besteht bei dieser Interpretation in der *semantischen Lücke* zwischen der Anforderungsspezifikation und der Darstellung der Lernmusterdaten, bzw. im der mangelnden Möglichkeit der Formalisierung der Semantik der Lernmusterdaten. Damit bleibt der Weg einer unmittelbaren formalen Verifikation verschlossen. Diese ist jedoch für hochsicherheitskritische Systeme empfehlenswert und wird von verschiedenen Normen für höhere Sicherheitsanforderungsstufen (*safety integrity level*, SIL, auch *Sicherheitsstufe* oder *Sicherheitsintegritätslevel*, vgl. [IEC02]) gefordert oder zumindest nahe gelegt (z. B. [ISO18]).

In jeder der beiden Interpretationen muss man für einen Korrektheitsnachweis des Systems einen Nachweis der Korrektheit der Lernmusterdaten führen. Dabei müssen die Daten in ihrer Gesamtheit mindestens die folgenden Eigenschaften erfüllen:

- **Relevanz:** Die Lernmusterdaten sollen den beabsichtigten Sachverhalt widerspiegeln
- **Vollständigkeit:** Die Lernmusterdaten sollen alle in diesem Sachverhalt möglichen Fälle (direkt oder indirekt) enthalten

- **Ausschließlichkeit:** Die Lernmusterdaten sollten keine nichtintendierten Relationen enthalten

Während die erste der geforderten Eigenschaften offensichtlich ist, bedürfen die anderen beiden ggf. einer Erläuterung:

Bei einer unendlichen Menge von möglichen Eingabewerten, wie dies bei fast allen nichttrivialen Anwendungen der Fall ist, können natürlich in den Lernmusterdaten eben nicht alle Fälle vorhanden sein. Dies ist jedoch wegen der Abstraktionsleistung des Netzes auch nicht notwendig, vgl. Abschnitt 3.2.3. Jedoch muss sichergestellt werden, dass die entscheidenden Grenzfälle richtig gelernt werden.

Die Ausschließlichkeit ist bei mehrdimensionalen Lernmusterdaten ein Problem: Neben der zu lernenden Information können andere, in der Anwendungspraxis unabhängige Informationen in der Menge der Lernmusterdaten enthalten sein. Dies ist bei komplexeren Anwendungen wie Realbildverarbeitung der Regelfall. Daher muss darauf geachtet werden, dass nicht in der Praxis unkorrelierte Fakten in einer (starken) Korrelation im Satz der Lernmusterdaten stehen, da sonst falsche Assoziationen gebildet werden.

4.3. Einfluss der Lernmusterdaten auf Verlässlichkeitsaspekte

Jenseits dieser speziellen Anforderungen hat die Einbeziehung von Lernmusterdaten in den Softwareentwicklungsprozess Auswirkungen auf verschiedene Aspekte der Verlässlichkeit (vgl. Abschnitt 2), die ein anderes Herangehen als in den anderen Systembereichen (vgl. Abschnitt 4.1) erfordern.

Die Zuverlässigkeit (*reliability*) wird nur insofern beeinträchtigt, dass inkorrekte Lernmusterdaten *latente* Fehler einbringen können. Bei den oben gegebene Interpretationen der Lernmusterdaten entspricht dies Entwurfsfehlern (*design faults*), die nichts prinzipiell Neues bei komplexer Software darstellen. Ähnliches gilt für die Verfügbarkeit: Im Endeffekt ist man bei beiden Aspekten auf die Korrektheit zurückgeworfen. Alle anderen eingehenden Faktoren unterscheiden sich ebenfalls nicht von anderen Systemen.

Die Instandhaltbarkeit (*maintainability*) ist durch den Einbezug von Lernmusterdaten in den Softwareentwicklungsprozess eingeschränkt: Fehler(zustände) in Lernmusterdaten sind im fertigen System nur schwer (wenn überhaupt) zu korrigieren. Jedoch kann unter Umständen das erneute Lernen mit korrekten Lernmusterdaten weniger aufwendig sein als das klassische Patchen von Software, da es eine klarere Schnittstelle gibt.

Die *Integrität* (*integrity*) eines KI-Systems ist außerhalb des Bereichs der Informationssicherheit durch den Einsatz von Künstlichen Intelligenz nicht unmittelbar betroffen, wenn die einmal erlernte Konfiguration während der aktiven Anwendungszeit eingefroren ist. In diesem Fall sind die üblichen Maßnahmen (z.B. kryptografische Signatur) zur Absicherung von Datenintegrität möglich. Allerdings sind aus der Perspektive der Informationssicherheit (*security*) durch Einsatz von Verfahren des maschinellen Lernens neue Angriffsarten möglich, die die Integrität und die Verfügbarkeit (Denial-of-Service-Angriffe) beeinflussen können, siehe Arbeitspaket V. Die *Vertraulichkeit* ist

durch Einsatz von Künstlichen Intelligenz nicht speziell betroffen. Freilich können durch die gesteigerte Komplexität neue verstecke Ausgabekanäle entstehen, was aber nicht KI-spezifisch ist.

Der Aspekt der technischen Sicherheit ist – wie oben bereits diskutiert – in erster Linie durch den Einsatz determiniert (der nicht KI-spezifisch), sowie durch Zuverlässigkeit und Verfügbarkeit, für die das oben Diskutierte gilt.

Die Tabelle 4.1 listet die Besonderheiten im Bezug zu den anderen übergeordneten Systemdomainen noch einmal kompakt auf.

Tabelle 4.1.: Besonderheit bei Verlässlichkeitsaspekten von sicherheitskritischen ML-Systemen in Abgrenzung zu anderen übergeordneten Systemdomänen

Aspekt	**Abgrenzung gegenüber**	**Unterschied**
Reliability/ **Availability**/ **Korrektheit**	klassische Systeme	Erforderliche semantische Korrektheit und Vollständigkeit von Lerndaten
Maintainability	klassische Software	Patching von Netzkonfiguration und/oder Lernmusterdaten schwer möglich
Fehlermodelle	klassische Systeme	Semantische Beschreibung von Fehlern in Lernmusterdaten fehlt weitgehend
Spezifikation	sichere Systemen	z. T. unscharfe Spezifikation auf Verhaltensebene
Implementation	klassische Software (außer MDD u.ä.)	Daten Teil des Implementationsprozesses
Testen	klassische Software	Ansätze zur Testung von Daten sind Gegenstand aktueller Forschung
formale Verifikation	sichere Software	Modelle zur Beschreibung semantischer Korrektheit sind Gegenstand aktueller Forschung

5. Thesen

An dieser Stelle werden wichtige Thesen dieses Reports kompakt zusammengefasst.

1. Systeme der Künstlichen Intelligenz sind **nicht ausschließlich** Systeme des maschinellen Lernens, Systeme des maschinellen Lernens des sind nicht ausschließlich neuronale Netze.

2. Nichtlernende Systeme der Künstlichen Intelligenz können bei der Nachweisführung **analog** den klassischen IT-Systemen der Anwendungsdomän behandelt werden. Dies gilt weitgehend auch für Systeme des maschinellen Lernens, sowie die Lernmusterdaten eine klare Semantik besitzen.

3. Neuroinspirierte KI-Systeme unterscheiden sich von anderen komplexen IT-Systemen bezüglich der technischen Sicherheit **nahezu ausschließlich** durch die Rolle der Lernmusterdaten.

4. Der Lernvorgang beim Systemen des maschinellen Lernens ist als Teil der Softwareentwicklung zu betrachten und nimmt die Rolle der der **Modellübersetzung** in der modellgeetriebenen Softwareentwicklung oder des **Codierens** in der klassischen Programmierung ein.

5. Das größte Hindernis zur überführen klassischer Ansätze und Verfahren zur Sicherstellung der Verlässlichkeit von Systemen mit neuronalen Netzen ist der **Mangel an semantischen Modellen** von Lernmusterdaten.

6. Derzeit können die geforderten Eigenschaften der Lernmusterdaten fast ausschließlich auf dem Weg von **Audits** und nur teilweise durch Testen verifiziert werden.

Teil II.

Glossar

Wichtige Begriffe und Abkürzungen

Accountability (*engl.*)

siehe Zurechenbarkeit

AI (*Abk.*)

Artificial Intelligence (*engl.*)

Aktivierungsfunktion

Funktion eines künstlichen Neurons, die aus dem Eingangsskalar und dem alten Aktivierungszustand einen neuen Aktivierungszustand bestimmt S.15

Aktivierungszustand

Zustand eines künstlichen Neurons, der nach dem Erregungszustand des biologischen Vorbilds modelliert ist S.15

Anforderungsspezifikation

Spezifikation, die beschreibt, *was* das System leisten soll, im Gegensatz zur Verhaltensspezifikation, die das Verhalten beschreibt, also *wie* das System dies leisten soll S.21

Artificial Intelligence (*engl.*)

siehe Künstliche Intelligenz

Auditieren

Untersuchung der Implementation eines Systems auf Schwachstellen S.10

Ausfall

Ereignis, an dem ein Dienst beginnt von der Erwartung abzuweichen S.8

Ausgabefunktion

Funktion eines künstlichen Neurons, die aus dem Aktivierungszustand eines künstlichen Neurons die Werte berechnet, die als Ausgabewertevektor zurückgegeben werden S.15

Availability (*engl.*)

siehe Verfügbarkeit [Attribut]

Bestärkendes Lernen

Trainieren eines künstliche neuronale Netzes mit Lernmusterdaten und Bewertung der Reaktion S.17

CIA-Triade

Vertraulichkeit, Integrität und Verfügbarkeit S.6

Confidentiality (*engl.*)

siehe Vertraulichkeit

Deep Learning (*engl.*)

siehe Tiefes Lernen

Dependability (*engl.*)

siehe Verlässlichkeit

Dienst eines Systems

Verhalten des Systems gegenüber einem Nutzer S.5

Entscheidungsbaum-Lernen

Methode der KI, bei der ein Entscheidungsbaum durch maschinelles Lernen parameterisiert wird S.14

Error (*engl.*)

siehe Fehlerzustand

Failure (*engl.*)

siehe Ausfall

Fault (*engl.*)

siehe Fehlerursache

Fehlertolerantes System

System, welches auch bei gewissen Fehlern seinen Dienst erbringt S.4

Fehlerursache

tatsächliche oder angenommene Ursache eines Fehlerzustands S.8

Fehlerzustand

Teil des Gesamtzustandes, der bei korrektem Verhalten nicht auftritt. S.8

Functional Safety (*engl.*)

siehe Funktionale Sicherheit

Funktionale Sicherheit

Abwesenheit unangemesser Risiken durch Gefahren wegen Fehlfunktionen S.20

Informationssicherheit

Verlässlichkeitseigenschaften mit Bezug zur Nutzerbefugnis S.6

Integrität

Verhinderung von unzulässigen Zustandsänderungen S.6

Integrity (*engl.*)

siehe Integrität

KI (*Abk.*)

Künstliche Intelligenz

Klassische Lernverfahren

Menge von Methoden des maschinellen Lernens, die eine (relativ) einfache Nachprüfbarkeit des Lernergebnisses gestatten S.13

Korrektheit eines Dienstes

Dienst verhält sich entsprechend der Erwartung S.8

Künstliche Intelligenz

Einheit von Hard- und Softwarekomponenten, welche – inspiriert von biologischen Vorbildern – zum Zweck der autonomen, effizienten und kreativen Problemlösung konzipiert, konstruiert und eingesetzt werden. S.12

Künstliche Intelligenz, schwache

KI zur Bewältigung konkreter einzelner Anwendungsproblemen des menschlichen Denkens bzw. der menschlichen Kognition S.12

Künstliche Intelligenz, starke

Nachbildung der menschlichen Intelligenz S.12

Machine learning (*engl.*)

siehe Maschinelles Lernen

Maintainability (*engl.*)

siehe Wartbarkeit

Maschinelles Lernen

Klasse von KI-Verfahren, bei denen Wissen aus Lernmusterdaten gewonnen wird; dies schließt künstliche neuronale Netze ein, aber auch weitere Verfahren . S.13

MDD (*Abk.*)

Model-driven development (*engl.*)

Mean Time To Failure (*engl.*)

siehe Mittlere Zeit bis zum Ausfall

Mittlere Zeit bis zum Ausfall

Erwartungswert der Zeit, bis ein Ausfall eintritt S.9

ML (*Abk.*)

Maschinelles Lernen

Model-driven development (*engl.*)

siehe Modellgetriebenen Entwicklung

Modellgetriebenen Entwicklung

Ansatz des Softwareengineerings, bei dem aus formalen Modellen, die typischerweise in einer anwendungsdomännahen Sprache beschrieben sind, automatisch lauffähige Programme generiert werden. S.20

MTTF (*Abk.*)

Mittlere Zeit bis zum Ausfall

Neuroinspirierte Verfahren

Klasse von Methoden der KI, die nach dem Vorbild der natürlichen (biologischen) Nervensysteme geschaffen wurden; künstliche neuronale Netze S.14

Nichtnutzer

Nutzer, der bezüglich eines bestimmten Systemdiestes keine Befugnis hat; unbefugter Nutzer S.6

Nutzer eines Systems

jeder Mensch oder jedes andere System, der/das mit dem System in Wechselwirkung steht S.5

Propagierungsfunktion

Funktion eines künstlichen Neurons, die den Eingangssignalvektor in einen Skalarwert wandelt S.15

Regelbasiertes Verfahren

Verfahren der Künstlichen Intelligenz, bei dem Aussagen aus (symbolischen) Daten (Fakten) und Schlussregeln gewonnen werden S.13

Reliability (*engl.*) [Attribut]

siehe Zuverlässigkeit [Attribut]

Reliability (*engl.*) [Maß]

siehe Zuverlässigkeit [Maß]

Robustheit

Tolierierung inkorrekter Eingaben/Umweltbedingungen S.7

Robustness (*engl.*)

siehe Robustheit

Safety (*engl.*)

siehe Technische Sicherheit

Safety Integrity Level (*engl.*)

siehe Sicherheitsanforderungsstufe

Security (*engl.*)

siehe Informationssicherheit

Service (*engl.*)

siehe Dienst eines Systems

Sicherheitsanforderungsstufe

Konzept aus dem Gebiet der funktionalen Sicherheit; eine Sicherheitsanforderungsstufe dient zur Beurteilung der Anforderungen an IT-Systeme im Bezug auf technische Sicherheit; einzelne Sicherheitsanforderungsstufen verlangen die Befolgung bestimmter Konstruktionsprinzipien oder die Einhaltung bestimmter Verlässlichkeitskennzahlen S.21

Sicherheitsintegritätslevel

siehe Sicherheitsanforderungsstufe

Sicherheitskritisches System

System, bei dem die Folgekosten eines Systemversagens mindestens eine Größenordnung höher sind, als die Kosten des eigentlichen Systems S.4

Sicherheitsstufe

siehe Sicherheitsanforderungsstufe

SIL (*Abk.*)

Safety Integrity Level (*engl.*)

System

aus mehreren Einzelteilen zusammengesetztes Ganzes S.4

Technische Sicherheit

Verhinderung von Schaden S.6

Testen

Überprüfung, ob die Reaktion des Systems auf ausgewählte Eingabedaten den in der Spezifikation des System angegebenen Verhalten entspricht S.10

Testen, erschöpfendes

Testen mit allen möglichen Eingaben; neben der formalen Verifikation der einzige Ansatz zum vollständigen Korrektheitsnachweis S.10

Tiefes Lernen

Lernansätze mit künstliche neuronale Netze, die mehr als eine Schicht von Neuronen zwischen Eingabe und Ausgabe aufweisen S.15

Trainieren eines neuronalen Netzwerks

Instanziierung (Ausprägung der Strukturen/Parameter) eines neuronalen Netzwerks durch eine Menge von Lernmusterdaten S.16

Überlebenswahrscheinlichkeit

siehe Zuverlässigkeit [Maß]

Überwachtes Lernen

Trainieren eines künstlichen neuralen Netzes mit Lernmusterdaten und zugehörigen Zielvorgaben S.16

Unüberwachtes Lernen

Trainieren eines künstlichen neuralen Netzes ohne Lernhilfen oder Bewertungsvorgaben nur mit Lernmusterdaten S.16

User (*engl.*)

siehe Nutzer eines Systems

Verfügbarkeit [Attribut]

Bereitstellung eines (nützlichen) Dienstes S.6

Verhaltensspezifikation

Spezifikation, die beschreibt, wie das System sich verhalten soll, mitunter auch als Spezifikation der Implementation, die das dann z. T. Verhalten indirekt beschreibt S.21

Verification (*engl.*)

siehe Verifikation

Verifikation

Nachweis des korrekten Verhaltens eines Systems S.9

Verifikation, formale

Nachweis des korrekten Verhaltens eines Systems mit Hilfe formaler Methoden; neben erschöpfendem Testen der einzige Ansatz zum vollständigen Korrektheitsnachweis S.10

Verlässlichkeit

Vertrauenswürdigkeit eines Computersystems, dass ein begründetes Vertrauen in den Dienst den es erbringt gesetzt werden kann. S.4

Vertraulichkeit

Verhindern von unbefugter Nutzung/Informationsoffenlegung S.6

Wartbarkeit

Möglichkeit zu Modifikationen und Reperaturen S.6

Zurechenbarkeit

Zurückverfolgbarkeit von Urhebern von Aktionen S.7

Zustand eines verlässlichen Systems

Berechnungen, Kommunikation, gespeicherte Information, Verbindungen und physische Gegebenheiten. S.4

Zuverlässigkeit [Attribut]

Kontinuität des Dienstes S.6

Zuverlässigkeit [Maß]

Wahrscheinlichkeit, dass ein System über einen Zeitraum $[t_0, t_0 + t]$ ein korrektes Verhalten zeigt, wenn es dies bereits zum Zeitpunkt t_0 tat. S.9

Literatur- und Quellenverzeichnis

[AL86] A. Avizienis and J.-C. Laprie. Dependable computing: From concepts to design diversity. *Proceedings of the IEEE*, 74(5):629–638, 1986.

[ALR+01] Algirdas Avižienis, Jean-Claude Laprie, Brian Randell, et al. Fundamental concepts of dependability. Technical report, University of Newcastle upon Tyne, Computing Science, 2001.

[ALR04] Algirdas Avižienis, Jean-Claude Laprie, and Brian Randell. Dependability and its threats: A taxonomy. In *Building the Information Society*, pages 91–120. Springer, 2004.

[ALRL04] A. Avizienis, J. C. Laprie, B. Randell, and C. Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing*, 1(1):11–33, January 2004.

[BBF03] Selmer Bringsjord, Paul Bello, and David. Ferrucci. Creativity, the turing test, and the (better) lovelace test. In James H. Moor, editor, *The Turing Test*, volume 30 of *Studies in Cognitive Systems*. Kluwer Academic Publishers, Dordrecht, 2003.

[Duv19] Alexandre Duval. Explainable Artificial Intelligence (XAI). 2019.

[HM19] Habib Hadj-Mabrouk. Contribution of Artificial Intelligence to Risk Assessment of Railway Accidents. *Urban Rail Transit*, 5(2):104–122, 2019.

[HWL17] Maren Henzel, Hermann Winner, and Benedikt Lattke. Herausforderungen in der Absicherung von Fahrerassistenzsystemen bei der Benutzung maschinell gelernter und lernender Algorithmen. 2017.

[ICE00] ICE/DIN. DIN EN 50126:1999 Bahnanwendungen – Spezifikation und Nachweis der Zuverlässigkeit, Verfügbarkeit, Instandhaltbarkeit und Sicherheit (rams), 2000.

[IEC02] IEC. IEC 61508:2001 functional safety of electrical/electronic/programmable electronic safety-related systems (german version en 61508:2001), 11 2002.

[ISO18] ISO. ISO 26262:2018 road vehicles – functional safety, 08 2018.

[J.L14] Hector J.Levesque. On our best behaviour. *Artificial Intelligence*, 217:27–35, 2014.

[Kri07] David Kriesel. *Ein kleiner Überblick über neuronale Netze*. 2007. http://www.dkriesel.com/science/neural_networks.

[KTAA07] Zeshan Kurd, A E Tim, Kelly Ae, and Jim Austin. Developing artificial neural networks for safety critical systems. 2007.

[Kur90] Raymond Kurzweil. *The Age of Intelligent Machines*. MIT Press, 1990.

[KW16] Philip Koopman and Michael Wagner. Challenges in Autonomous Vehicle Testing and Validation. 2016.

[Lan19] Holger Lange. Künstliche Intelligenz in Sicherheitssystemen, 2019.

[Lap85] Jean-Claude Laprie. Dependable computing and fault-tolerance. *Digest of Papers FTCS-15*, pages 2–11, 1985.

[Lap92] Jean-Claude Laprie. *Dependability: Basic concepts and terminology*. Springer, 1992.

[Lap95] Jean-Claude Laprie. Dependable Computing: Concepts, Limits, Challenges. In *Special Issue of the 25th International Symposium on Fault-Tolerant Computing*, pages 42–54, Pasadena, California, USA, 1995.

[LK96] Jean-Claude Laprie and Karama Kanoun. Software reliability and system reliability. page 45, 1996.

[MRSM06] Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon, and John McCarthy. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. 27(4):12, 2006.

[Nil98] Nils J. Nilsson. *Artificial Intelligence – A New Synthesis*. Morgan Kaufmann, 1998.

[Ric83] Elaine Rich. *Artificial Intelligence*. McGraw-Hill, 1983.

[Rod02] David M Rodvold. A Software Development Process Model for Artificial Neural Networks in Critical Applications. 2002.

[SL10] Johann Schumann and Yan Liu. *Applications of Neural Networks in High Assurance Systems: A Survey*. Springer, Berlin, Heidelberg, 2010.

[SQC17] Rick Salay, Rodrigo Queiroz, and Krzysztof Czarnecki. An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software. 2017.

[Tay06] Brian J. Taylor. *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*. Springer US, 1 edition, 2006.

[Tur50] Alan Mathison Turing. Computing machinery and intelligence. *Mind*, LIX(236):433—-460, 1950.

[WPW+18] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient Formal Safety Analysis of Neural Networks. Technical report, 2018.

Beschreibung der gewählten bahnspezifischen Beispiele

Zweck und Inhalt des Dokuments

Das vorliegende Dokument dient der Dokumentation der im Rahmen des AP4 „Reflexion an bahnspezifischen Beispielen“ des Projektes „KI-bezogene Test- und Zulassungsmethoden“ gewählten Beispiele. Die Auswahl wurde beim virtuellen Projekttreffen am 9.12.20 vorgestellt und diskutiert und im Folgenden genauer ausgearbeitet. Die aktuelle Version des Dokuments beinhaltet minimale Korrekturen nach der Vorstellung und Diskussion der ausgearbeiteten Beispiele beim virtuellen Treffen am 4.2.21. Zudem wurden im Rahmen der Bearbeitung von AP5/6 (virtueller Workshop am 7.12.21 und nachfolgend durch IVS erstellter Nachweisansatz) Konkretisierungen der Beispiele 1 und 3 nötig, die in das vorliegende Dokument aufgenommen wurden.

Die Beispiele sollen zur Reflexion der Projektergebnisse weiterer APs an konkreten Anwendungsfällen aus dem Bahnbereich dienen. Dies betrifft insbesondere, aber nicht ausschließlich, die Ergebnisse zum rechtlichen Rahmen sicherheitsrelevanter Bahnanwendungen (AP2) sowie zur technischen Sicherheit von KI (AP3).

Falls es im Projektverlauf Gründe dafür gibt, können die Beschreibungen der Beispiele in diesem Dokument in einem gewissen Rahmen ergänzt oder angepasst werden. Entscheidend ist, dass sie für die Projektpartner nutzbar sowie realistisch und in sich kohärent bleiben.

In einem ersten Teil des Dokuments wird die Auswahl der Beispiele kurz beschrieben. In den folgenden Teilen werden die drei gewählten Beispiele jeweils mit ihrem Hintergrund, ihrer Definition, ggf. Überlegungen zu Systemarchitektur und zuverlässigkeitssteigernden Maßnahmen, und Charakteristika bzgl. Sicherheit und KI vorgestellt. Zitierte Literatur ist am Ende des Dokuments zusammengestellt.

Auswahl der Beispiele

Auswahlkriterien

Die zu wählenden Beispiele sollten, um dem Ziel des Projektes bestmöglich zu dienen, die folgenden Kriterien erfüllen:

- Bahnspezifisch
- KI-Anwendung nötig oder möglich
- Sicherheitsrelevanz (verschiedene Stufen)
- Klar erkennbarer Nutzen für Stakeholder im Bahnsystem
- Konkret genug und dennoch breit genug für Anknüpfung an verschiedene
    - KI-Ausprägungen (AP 1/3)
    - Rechtskontexte (AP 2)
    - Safety-Mechanismen (AP 3)
- Vergleichspunkte mit anderen Domänen (AP 1)
- Technische Umsetzung realistisch (ab TRL 3/4)
- Möglichst Publikation(en) verfügbar

Beispielkandidaten

Es wurden zahlreiche Beispiele für die Nutzung von KI im Bahnbereich ermittelt. Diese besitzen sehr unterschiedliche Reifegrade und Sicherheitsrelevanz ihrer Anwendungen. Die Beispiele wurden in Gruppen sortiert:

1. Objekterkennung im Betrieb aus Bilddaten oder Punktwolken
    a. Hindernisse auf/an der Strecke
    b. Landmarken
    c. Signalbegriffe
    d. Weichenlage
    e. Passagiere (Anzahl)
    f. Verdächtige Gegenstände/Personen
2. Zustandsüberwachung, -diagnose und -prädiktion anhand von Sensordaten
    a. Oberbau (Gleise, Weichen)
    b. Kabelanlagen
    c. Achsen von Zügen
3. Zustandsüberwachung per Analyse von (Drohnen-)Bildern
    a. Schäden an Oberleitungsisolatoren
    b. Schäden an Bauten (Gebäude, Brücken)
    c. Ausmaß Vegetation/Grünbewuchs
    d. Vorhandensein diebstahlgefährdeter Metallteile (Kabel, Oberleitung)
    e. Risse in Schwellen
4. Erkennung von Bewegungen auf/an der Strecke aus Fiber Optic Sensing Daten
    a. Hindernisse (Tiere, Steinschlag)
    b. Metalldiebe
    c. Züge
    d. Räder (Laufverhalten auf den Schienen)
5. Verkehrsanalyse und -optimierung basierend auf Verkehrsdaten

a. Prädiktion des Verkehrsgeschehens (Ankunftszeiten)
b. Ursachenanalyse bei Planabweichungen
c. Dispositionsentscheidungen
d. Fahrassistenz (Optimalgeschwindigkeit)

Daneben wurde festgestellt, dass in vielen sicherheitsrelevanten Bahn-Kontexten die Anwendung von KI zwar technisch möglich, jedoch nicht sinnvoll ist. Dies umfasst Beispiele, in denen ausreichend effiziente analytische Lösungen existieren sowie Fälle, in denen es quantitativ oder qualitativ an geeigneten Trainingsdaten für überwachte maschinelle Lernverfahren fehlt und in denen solche Daten schwer zu erzeugen sind.

Gewählte Beispiele

Aus Mangel an Sicherheitsrelevanz wurden die Beispiele aus Gruppe 5 (Verkehrsanalyse und -optimierung) ausgeschlossen; ebenso wurden Beispiele mit Security-Hintergrund (1f, 3d, 4b) aussortiert, da der Projektkontext eine Fokussierung auf Safety (Unfallvermeidung) nahelegt, wo die im Bahnbereich etablierten Sicherheitsmaßnahmen eine Einführung KI-basierter Systeme erschweren. Schließlich wurden – unter Berücksichtigung der oben genannten Kriterien – folgende drei Beispiele aus den Bereichen 1a, 2a und 4c gewählt:

1. Hinderniserkennung auf/neben dem Gleis per Kamera
2. Zustandserfassung und Fehlerdiagnose für Gleise und Weichen per Multisensor-System
3. Bestimmung von Zugposition und -vollständigkeit per Fiber Optic Sensing

Sie stammen aus verschiedenen Gruppen (1, 2, 4) und decken damit eine gewissen Bandbreite ab. Die Beispiele aus Gruppe 3 weisen technisch Ähnlichkeiten mit denen aus Gruppe 1 auf (Bilddatenanalyse) und sind weniger sicherheitsrelevant als die ebenfalls wartungsbezogenen Beispiele aus Gruppe 2, so dass durch ihren Ausschluss mutmaßlich keine wesentlichen Aspekte verloren gehen. Die gewählten Beispiele qualifizieren sich auch durch ihr Potenzial, zu den aktuell viel diskutierten Themen des automatisierten Fahrens (Beispiel 1), der zustandsbasierten Instandhaltung (Beispiel 2) sowie des „Moving Block“ (Beispiel 3) beizutragen.

Beispiel 1: Hinderniserkennung auf/neben dem Gleis per Kamera

Hintergrund

Der Triebfahrzeugführer (Tf) ist durch die DB-Richtlinie 408 [1] Modul 2341 verpflichtet, während einer Zugfahrt die Strecke zu beobachten. Dies dient der Erkennung möglicher Gefährdungen für den eigenen Zug und/oder weitere Züge, die die Strecke nutzen. Zu solchen Gefährdungen zählen z. B.

- Personen im/am Gleis,
- größere Tiere im/am Gleis,
- Straßenfahrzeuge auf dem Gleis,
- umgestürzte Bäume, abgegangene Erdmassen/Steine/Geröll,
- nicht korrekt gesicherte Bahnübergänge,
- defekte Oberleitungen und
- defekte Signalanlagen.

Während das DB-Regelwerk bisher Zugfahrten ohne Tf nicht kennt, wird das automatisierte Fahren (ATO – automatic train operation) im Bahnbereich zunehmend thematisiert [2]. In den höheren Automatisierungsgraden (GoA – Grade of Automation, vgl. IEC 62267 [3]) 3 und 4 ist kein Tf mehr an Bord. Einer der herausforderndsten und vielschichtigsten Aspekte der Realisierung ist dabei, die Streckenbeobachtung des Tf adäquat/sicher durch technische Systeme zu ersetzen. Potenzielle Vorteile technischer Systeme umfassen

- ungeteilte Aufmerksamkeit/keine Ablenkung,
- verlässliche und nachprüfbare Situationserfassung,
- automatische und damit schnelle Meldung von Gefährdungen, und
- wählbare Perspektive/Installationsort;

potenzielle Stärken der Beobachtung durch den Menschen dagegen

- Antizipationsfähigkeit von Objekten aufgrund lediglich kleinster sichtbarer Ausschnitte oder Änderungen/Bewegungen,
- weites Sichtfeld/peripheres Sehen, und
- Nutzung von Erfahrung und Kontextwissen bei der Erkennung und Bewertung der Situation.

Definition

Beispiel 1 konzentriert sich auf die Erkennung von Hindernissen auf oder neben dem Gleis mit Hilfe einer oder mehrerer auf dem Zug installierter Kameras. Diese zeichnen in regelmäßigen zeitlichen Abständen Bilder auf bzw. nehmen ein Video auf. Die Bilder bzw. das Video beinhalten einen Ausschnitt einer Bahnstrecke mit einem oder mehreren Gleisen und deren unmittelbarer Umgebung.

Zur Erkennung und Klassifizierung von Hindernissen sowie der Bestimmung ihrer gleisgenauen Position, der Entfernung vom Zug und der Hindernisgröße[1] wird fortlaufend eine KI auf die Bilder angewandt, deren Ergebnisse jeweils spätestens einige Sekunden nach Aufzeichnung des

[1] Die KI selbst leistet klassischerweise die Objekterkennung (Identifikation von Bildregionen mit Objekten und Klassifizierung der Objektart). Zur Bestimmung der 3D-Position und -Größe des Objekts wird ein nachgelagertes (nicht-KI) System angenommen, das neben der durch die KI bestimmten Bounding Box weitere Informationen (z. B. Radar) als Eingabe haben kann.

ausgewerteten Bildes vorliegen. Aus Datenschutzgründen und aus Gründen des Datenvolumens erfolgt diese Auswertung ohne weiträumige Datenübertragung lokal am Standort der Kamera. Eine zeitlich beschränkte Historie von Bildern oder Ergebnissen kann in die aktuellen Ergebnisse einfließen, beispielsweise um Ergebnisse zu plausibilisieren oder Trajektorien bewegter Hindernisse zu bilden / Bewegungsmuster zu erkennen.

Die KI stellt ihre Ergebnisse inklusive Unsicherheiten einem lokalen oder entfernten System zur Verfügung, das, ggf. unter Berücksichtigung weiterer Informationen wie Zugpositionen und -geschwindigkeiten, eine angemessene Reaktion auslöst. Dies kann die Einleitung einer sofortigen Not- oder Betriebsbremsung von Zügen sein, eine Meldung an Züge in der Umgebung, die Benachrichtigung von Betriebs-, Wartungs-, Rettungspersonal oder Behörden, das Abwarten, ob sich das Ergebnis manifestiert, oder das Ignorieren des Ergebnisses.

Ein entsprechendes System ist in Abbildung 1 skizziert. Je nach Betrachtungsfokus liegen alle vier rechteckigen Boxen innerhalb der Systemgrenzen (da sie gemeinsam den Tf ersetzen), lediglich die obere Box (da hier die KI zum Einsatz kommt), oder eine Teilmenge der Boxen (wenn man z. B. die Fragestellung der Situationsbewertung und Reaktion als separates Problem betrachtet).

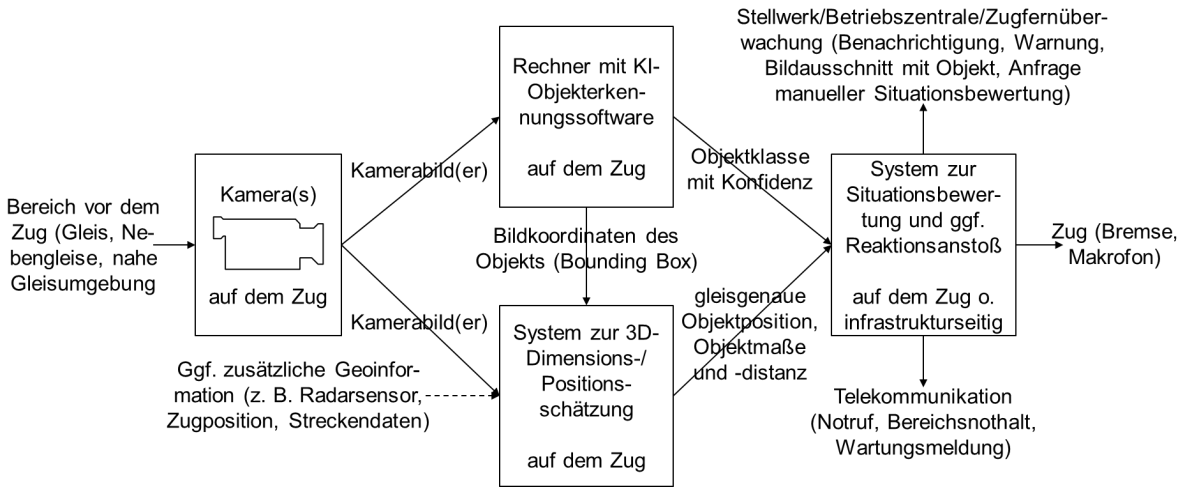


Abbildung 1: Skizze des Systems zur Hinderniserkennung

Beispiel 1 hat den Anspruch, die Streckenbeobachtung durch den Tf – ausgenommen die Beobachtung der bahntechnischen Anlagen – weitgehend zu ersetzen in dem Sinne, dass es möglicherweise für bestimmte Situationen einer Ergänzung durch andere technische Systeme bedarf (z. B. für Nachtsicht, für vereinzelte Ergänzung durch ortsfeste Beobachtung kritischer Punkte/Abschnitte an der Strecke, oder für die Bereitstellung von Kontextinformationen zum Kamerabild). Solche anderen Systeme dürfen als vorhanden und ausreichend zuverlässig angenommen werden, soweit dies realistisch erscheint und die zugseitigen Kamerabilder zentrale Grundlage für die Hinderniserkennung bleiben.

Beispiel 1 bezieht sich geografisch auf Deutschland und umfasst folgende Situationen:

- Verschiedene betriebliche Rahmenbedingungen (Voll- oder Nebenbahnen, Personen- oder Güterverkehre, Bahnen nach EBO oder BOStrab, etc.);
- verschiedene Arten von Hindernissen (siehe z. B. die ersten vier Anstriche der ersten Aufzählung im obigen Abschnitt „Hintergrund“ sowie unterhalb dieser Aufzählung);

- verschiedene Größen von Hindernissen;
- Hindernis auf dem Gleis und/oder am Gleis;
- verschiedene Streckenumgebungen (Wiese, Feld, Wald, Böschung, Straße, Häuser, Brücke, Tunnel, Bahnhof, etc.);
- unterschiedliche Tageszeiten und Wetterverhältnisse (von einer gewissen Grundausleuchtung darf ausgegangen werden, vgl. obige Anmerkung zu ergänzenden technischen Systemen).

Da die Definition der zu erkennenden Hindernisarten zentral für das System ist, wird an dieser Stelle eine mögliche Liste von Hindernisklassen und weiter zu differenzierenden Unterklassen skizziert (ohne den Anspruch, dass diese Klassifizierung für den Anwendungszweck notwendig oder hinreichend ist):

1. Personen (auch inkl. Fortbewegungsmittel wie Fahrrad, Kinderwagen, Rollstuhl, Roller, etc.)
    a. Kinder/Erwachsene
    b. Bahnpersonal/Unbefugte
    c. Warngeste (z. B. Kreissignal/Sh3)
2. Fahrzeuge
    a. Straßen-/Schienenfahrzeug
    b. Ggf. Motorrad/LKW/PKW/Andere
    c. Ggf. Fahrtrichtung Schienenfahrzeug (vorausfahrend/entgegenkommend)
3. Tiere
    a. Ggf. Nutztiere/Wildtiere
    b. Ggf. genauere Tierart (Auswahl relevanter Arten, z. B. Kuh, Pferd, Schaf, Hirsch, Reh, Wildschwein, Ziege, Hund, Schwein, Wolf)
4. Natürliche Objekte
    a. Bäume+Äste/Erdmassen+Steine+Geröll/Wassermassen/Schneemassen
5. Künstliche Objekte
    a. Auswahl relevanter Objekte, z. B. Müllcontainer, Schuppendächer, Betonteile, typische Schienentransportgüter wie Schüttgüter oder Holzstämme, etc.)
6. Brände

Dazu ist zu bemerken, dass, wie in Abbildung 1 dargestellt, die Hindernis-/Objektart nicht allein entscheidend für die Reaktion des Systems ist, sondern zusätzlich Position und Größe des Objekts hinzugezogen werden. Die ermöglicht wichtige Unterscheidungen wie z. B. ob Geröll sich auf dem eigenen oder Nachbargleis oder beiden befindet oder ob ein Brand ein Grillfeuer oder ein Böschungsbrand ist.

Aus sicherheitlicher Sicht können viele der Klassen bzw. Unterklassen möglicherweise wieder zusammengefasst werden. Entscheidend ist, welche unterschiedlichen Reaktionen für welche Klasse einzuleiten sind. Im einfachsten Fall ist lediglich zu entscheiden, ob eine (maximale) Sicherheitsreaktion (z. B. Notbremsung + Makrofon + Notruf + Information Fahrdienstleiter) eingeleitet werden muss oder nicht; es ist vorstellbar, dies anhand der Objektgröße zu entscheiden, indem eine Grenze definiert wird, unterhalb der ein Zusammenprall normalerweise keine ernsthaften Schäden verursacht, weil das Objekt keine Menschen beinhaltet und seine Masse gering ist. Hierfür wäre eine KI möglicherweise nicht einmal nötig. Aus verschiedensten Gründen kann es jedoch sinnvoll sein, die Reaktionen feiner abzustufen bzw. zusätzliche Reaktionen zu definieren. Solche Gründe können das Vermeiden von ressourcenintensiven Fehlalarmen von Rettungsdiensten, das Ermöglichen eines

möglichst unterbrechungsfreien Bahnbetriebs, die Realisierung von schnellen/zielgenauen Reaktionen oder eine differenzierte statistische Erfassung der Objekte sein. Dabei ist jeweils darauf zu achten, dass in jedem Fall eine ausreichende Sicherheit immer noch gewährleistet ist, wobei die Risiken eines möglichen Zusammenpralls für Zug, Objekt und Umgebung zu bewerten sind. Interessante Unterscheidungen neben der Größe, der Position auf oder neben dem Gleis und neben den oben bereits aufgelisteten Klassen könnten auf diesem Hintergrund die Bewegung bzw. Ausbreitung des Objekts (ja/nein, welche Richtung), das Vorhandensein von Menschen im Objekt, die Massivheit des Objekts oder die Klassifizierung nach Zuständigkeit eines bestimmten Rettungs- oder Wartungsdienstes sein. Auch kann es interessant sein, Korrelationen zu nutzen, um Größe oder Position indirekt/zusätzlich über die KI zu beurteilen (z. B. kleine Tierarten vs. große Tierarten, Objekt mit Schiene darunter vs. Objekt mit Schiene daneben).

Eine genauere Definition/Einschränkung der Systemumgebung für eine erste Version eines Systems könnte wie folgt aussehen (die angegebenen konkreten Werte basieren auf Bahnwissen, Recherchen und Überlegungen des Autors; es wird keine Gewähr für die Eignung für reale Systementwicklungen übernommen):

- Zug
    - Maximalgeschwindigkeit 160 km/h
    - Bremsvermögen eines Regionaltriebzugs (ca. 0,9 m/s²)
- Bahninfrastruktur
    - Kurvenradien von mindestens 2000 m (dies entspricht einem maximalen Winkel zwischen der Sichtachse zu einem 300 m entfernten Punkt auf dem Gleis und der geraden Sichtachse von etwa 4,3° -> unter Berücksichtigung der gewünschten Erfassung von Nachbargleisen sowie Links-/Rechtskurven ergibt sich ein zu erfassender Winkel von etwa 10°)
    - Maximale Steigung 2,5% (insbesondere keine Kuppen mit beidseitig höheren Steigungen)
    - Alle Bahnübergänge sind technisch gesichert, und bei bisher lokführerüberwachten Bahnübergängen wird die Information über die korrekt erfolgte Sicherung durch ein separates System erfasst
    - Bahnsteige sind mit einer separaten, lokalen Gefahrenerkennung ausgestattet
- Lichtverhältnisse & Witterungsbedingungen
    - Tageslicht
    - Beliebiger Sonnenstand und verkehrsübliche Lichter (Zugspitzensignal entgegenkommender Züge, Frontscheinwerfer von KFZ, etc.)
    - Sichtweiten von mindestens 300 m bzw. verringerte Sichtweiten nur bei reduzierter Geschwindigkeit
- Verdeckung
    - Vegetation in den Rückschnittzonen (6 m von Gleismitte nach links und rechts) maximal 50 cm hoch (bzw. nur vereinzelt höher)
    - Keine Bauten/Schilder/etc. in nicht zutrittsgeschützten Bereichen 6 m von Gleismitte nach links und rechts, die kumuliert (inkl. der sich möglicherweise darunter befindenden Vegetation) mehr als 0,5 m² zusammenhängende Flächenverdeckung

verursachen (ansonsten Maßnahmen zum Zutrittsschutz oder separate, lokale Überwachung)

Weitere Umgebungsbedingungen können je nach Positionierung und Befestigung des Systems auf dem Zug mehr oder weniger relevant sein; ihre Beachtung ist jedoch standardmäßig für Komponenten auf Bahnfahrzeugen geboten, weswegen hier von einer genaueren Definition abgesehen wird. Hierzu zählen der Umgebungstemperaturbereich, Windgeschwindigkeiten, Erschütterung/Schläge, Lateralkräfte und Vibrationen, Steinschlag und Kollisionen mit z. B. Vögeln außen am Fahrzeug, Staub/Schmutz/Insekten/Feuchtigkeit entsprechend des dauerhaften Einsatzes in Außenbereichen bzw. speziell auf Bahnstrecken und in Bahnbetriebsvorrichtungen, und elektromagnetische Bedingungen für elektronische Geräte auf dem Zug [18, 19].

Das Beispiel lässt offen (Festlegung technischer Details nur bei Bedarf und Konsens der Projektpartner),

- wie viele Kameras genutzt werden,
- welche Bildauflösung die Kameras liefern,
- welche Bildrate geliefert wird bzw. wie viele Bilder per KI ausgewertet werden,
- wie die Kamera(s) verbaut wird/werden, und
- wie die Kameralinsen sauber gehalten werden.

Nicht unter Beispiel 1 fallen:

- Rangierfahrten (hier hat die Pflicht zur Streckenbeobachtung abweichende Hintergründe, s. DB-Richtlinie 408 [1] Modul 4814);
- die Erkennung von Objekten, die keine Gefährdung für den Zug oder Menschenleben darstellen (Blätter und kleine Zweige, Insekten, Vögel und Kleintiere, Haushaltsabfälle, etc.);
- die Erkennung von Unregelmäßigkeiten und Defekten an Bahnanlagen (siehe z. B. die letzten drei Anstriche der ersten Aufzählung im obigen Abschnitt „Hintergrund“);
- die genauere Definition der Nutzung von Daten anderer Sensoren wie Laserscanner, Radar, Lidar, etc., die detaillierte zusätzliche Informationen zum Kamerabild liefern.

Um möglichst vollständige Anforderungen für ein Objekterkennungssystem zu formulieren, sind eher allgemeine Beschreibungen von Objekten (z. B. „Mensch“, „heimisches Tier > 0,8 m“, „Straßenfahrzeug“, „Schienenfahrzeug“) und Umgebungsparametern (z. B. „Streckenrandbebauung/-bewuchs“, „Oberbauform“, „Wetter“) sowie eine exakte Definition des mindestens zu erfassenden Bereichs („um die Schienenmitteltangente symmetrischer 10°-Winkel“, „bis 300 m Luftlinie von der Zugspitze entfernt“) nützlich. Um aus den allgemeinen Beschreibungen konkrete Anforderungen an geeignete Trainings- und Testdatensätze abzuleiten, sind Aussagen zu Mindestberücksichtigung, relativer Berücksichtigung (z. B. entsprechend des Vorkommens in der Realität), Diversität (z. B. „Menschen verschiedener Größe, Statur, Geschlecht und Herkunft“), Randfällen (z. B. „Objekt zu 10% oder 50% im erfassten Bereich“, „Mensch mit außergewöhnlicher Kleidung/Kostüm“, „Objekt vor Hintergrund mit geringem Kontrast“) und Kombinatorik (Kombinationsgrad von Objekten, Objektpositionen und Umgebungen auf verschiedenen Verfeinerungsebenen, mehrere ähnliche oder verschiedene Objekte im Erfassungsbereich) nötig. Um eine Überprüfbarkeit zu gewährleisten, sind die Anforderungen um Aussagen zu Konfidenzen zu ergänzen (im einfachsten Fall z. B. „alle Objekte müssen in jedem Testfall mit einer Konfidenz von mindestens 60% erkannt werden“). Zusätzlich sind

Anforderungen an die Robustheit der Objekterkennung zu formulieren (z. B. „Wenn kein Objekt da ist, bleiben alle Objektkonfidenzen unter 30%“, „Bei Menschen auf Werbeplakaten oder Verkehrsschildern bleibt die Konfidenz unter 30%“).

Überlegungen zu Systemarchitektur und zuverlässigkeitssteigernden Maßnahmen

Es ist wegen der Vielzahl möglicher Objekte, deren Varianten und Positionen sowie der Vielzahl möglicher Umgebungen, Wetter- und Beleuchtungsbedingungen davon auszugehen, dass die Erstellung der Trainingsdaten für die KI zur Objekterkennung ein sehr aufwändiger Prozess sein wird. Darüber hinaus wird ein Teil potenzieller Trainingsdaten für den Test der KI zurückgehalten werden müssen. Die Tests selbst können ebenfalls sehr (zeit)aufwändig werden. Wird davon ausgegangen, dass eine Zulassung nur für eine „eingefrorene“ trainierte und nachgewiesene KI erteilt wird, d.h. auch Änderungen entsprechend aufwändig sind, ist es von hoher Bedeutung, im ersten (vollumfänglichen) Versuch eines KI-Trainings und -Nachweises die geforderte Zuverlässigkeit zu erreichen. Zugleich scheinen übliche Objekterkennungsraten durch KI nicht ausreichend und die Objekterkennung kann mit unterschiedlichen Konfidenzen behaftet sein.

Um in dieser Situation KI überhaupt erfolgreich einsetzen zu können, sind zuverlässigkeitssteigernde Maßnahmen bei der Systemauslegung essenziell. Solche Maßnahmen können umfassen:

- Die Berücksichtigung der durch die KI mit ausgegebenen Konfidenzen. Beispielsweise kann ein Schwellwert definiert werden, unterhalb welchem die Objektklassifikationsaufgabe zurück an einen menschlichen Überwacher (in einer Betriebszentrale, in die das Bild weitergeleitet wird) delegiert wird.
- Ergebnisplausibilisierung. Anhand weiterer Daten, z. B. über die Streckenumgebung, kann die Objektklassifikation plausibilisiert und im Zweifelsfall wiederum einem menschlichen Überwacher zur Bestätigung vorgelegt werden (Reduktion der false positives). Beispielsweise ist es unwahrscheinlich, dass ein Baum auf den Schienen liegt, wenn die Strecke durch baumlose Felder und Wiesen führt, oder dass ein Erdrutsch auf einer Brücke stattgefunden hat.
- Diversitäre KI-Algorithmen und Strukturen neuronaler Netze. Ohne notwendigerweise mehr (disjunkte) Trainingsdaten zu benötigen, können unterschiedliche KI-Algorithmen und/oder Strukturen neuronaler Netze trainiert werden. Beispielsweise könnte eine KI sehr schnell Ergebnisse liefern und mit hoher Frequenz angewendet werden, und eine zweite nach längerer Rechenzeit genauere Resultate liefern und mit geringerer Frequenz bzw. bei Bedarf (z. B. unsicheren Ergebnissen der ersten KI) ausgeführt werden.
- Mehrkanaligkeit. Die Nutzung zweier Kameras und zweier dahinterliegender KI-Rechner erscheint eine kostengünstige und effektive Lösung zur Zuverlässigkeitssteigerung, da technisch im Wesentlichen kein zusätzlicher Entwicklungsaufwand anfällt (insbesondere ist keine diversitäre KI-Entwicklung nötig) und im Gegensatz zu den vorgenannten Maßnahmen eine Steigerung der Objekterkennungsrate erreicht werden kann (Reduktion der false negatives). Die erzeugten Eingabebilder für die KI und das System zur Dimensions-/Positionsschätzung sind dabei wegen der leicht unterschiedlichen Perspektive der Kameras unterschiedlich, wodurch neben der Ausfallsicherheit des Gesamtsystems die Zuverlässigkeit der Objekterkennung gesteigert wird und zudem eine einfachere/exaktere/allein kamerabildbasierte Berechnung der 3D-Objektposition ermöglicht wird.

- Modularisierung. Es ist denkbar, statt einer KI verschiedene, spezialisierte KIs zu trainieren (für verschiedene Objektklassen, für verschiedene Umgebungen oder Lichtverhältnisse). Auf diese Weise könnten genauere Ergebnisse und höhere Konfidenzen erzielt werden und notwendige Änderungen des Systems wären weniger aufwändig. Dabei ist zu beachten, dass sich möglicherweise Kosten und/oder die benötigte Rechenzeit des Systems erhöhen.
- Sequenzielle Verknüpfung. Da regelmäßig Bilder aufgenommen und auf Objekte analysiert werden, können die Ergebnisse über die Zeit verglichen werden, um so zu einer zuverlässigeren Aussage zu gelangen. Hierbei sind jedoch gegenläufige Effekte zu beachten: während der Zeitspanne nähert sich der Zug i.d.R. dem Objekt, was die Erkennung weiter verbessern kann; wenn jedoch mit einer Reaktion gewartet wird, steigt das Risiko einer Kollision mit dem Objekt. Object Tracking Techniken können genutzt werden, um die Bewegung eines Objektes zu verfolgen, was für die Robustheit des Systems genutzt werden kann (Vermeidung einer Notbremsung bei Fußgängern/Radfahrern, die die Schienen in einiger Entfernung vor dem Zug „noch schnell“ queren), oder aber die Objekterkennung auf zeitlich nachfolgenden Bildern effizienter zu machen (s. z. B. [16]).

Charakteristika

Sicherheitsrelevanz: Je nach Hindernis können sich keine bis katastrophale Folgen (Entgleisung) für den Zug ergeben. Befinden sich Personen im Gleis, sind auch für sie Folgen ganz verschiedenen Ausmaßes (keine bis hin zum Tod) möglich. Ebenfalls je nach Hindernis sehr unterschiedlich sind die durchschnittlichen Auftretenswahrscheinlichkeiten; daneben weisen sie starke Abhängigkeiten z. B. von Wetterlage, Örtlichkeit oder Tages-/Nachtzeit auf. Präventive Gegenmaßnahmen werden teils großflächig (z. B. Gehölzrückschnitt), teils lokal (z. B. Sicherung gegen Steinschlag, partielle Streckeneinzäunung) bereits getroffen. Systembedingt ergeben sich geschwindigkeitsabhängig Bremswege von bis zu mehreren Kilometern Länge, wodurch ein Anhalten vor einem erkannten Hindernis oft nicht mehr möglich ist (ggf. jedoch eine Geschwindigkeitsreduktion). Eine besondere Herausforderung für die Abwägung zwischen Sicherheit und Betrieb stellen Bahnhöfe/Haltepunkte als explizite Zugangspunkte für Personen zu Zügen dar.

Sicherheitsbetrachtung: Für die Streckenbeobachtung durch den Tf gibt es offensichtlich keine Quantifizierung der Anforderungen; deren Sinn wäre aus vielfältigen Gründen auch hinterfragbar. Stattdessen wird üblicherweise eine allgemeine Eignung im Rahmen der Tf-Ausbildung und in regelmäßigen Überprüfungen festgestellt sowie Streckenkenntnis vorausgesetzt, bevor ein Tf eine Strecke befahren darf. Entsprechend gibt es keine belastbaren Anforderungen für die bisher nicht vorgesehene Streckenbeobachtung durch technische Systeme. Essenziell für das Ermöglichen eines Sicherheitsnachweises solcher Systeme durch Vergleich mit dem bisherigen Verfahren wäre ein allgemein anerkannter Rahmen, der es erlaubt, einen statistisch-risikobasierten Vergleich für die Hinderniserkennung als Ganzes zu verwenden, da der Mensch technischen Systemen in Teilaspekten in absehbarer Zeit überlegen bleiben wird (s. Ende des Abschnitts „Hintergrund“ oben) bzw. eine entsprechende Ausrüstung mit technischen Systemen unverhältnismäßig teuer würde. Es ist zu berücksichtigen, dass die Akzeptanz von Risiken bei technischen Systemen allgemein geringer als die für menschliche Fehler ist. Noch hilfreicher wäre ein validiertes und akzeptiertes Regelwerk, das Mindestanforderungen in Form konkreter Werte wie zu erkennender Objektarten und -größen, zu betrachtender Umgebungsradien etc. sowie jeweiliger auszulösender Reaktionen beinhaltet. Teils könnten bereits in Betrieb befindliche automatische Metros/U-Bahnen (z. B.

Gleisbereichsüberwachung der fahrerlosen U-Bahn Nürnberg in Bahnhöfen) oder an Bahnübergängen eingesetzte Hinderniserkennungssysteme Anhaltspunkte hierfür liefern.

Gründe für KI-Nutzung: Für die Erkennung von Hindernissen gibt es keine klare Spezifikation. Das Bild einer einzigen Art von Hindernis (z. B. einer Kuh) variiert beispielsweise durch die Betrachtungsperspektive und -distanz, die Größe, Form, Helligkeit und Farbe des Objekts, sowie bei beweglichen Objekten durch die momentane Haltungsposition. Die unklare Spezifikation trifft nicht nur auf das Hindernis selbst, sondern ggf. auch auf weitere im Kamerabild enthaltene relevante Objekte und Anordnungen zu (z. B. einen Zaun oder eine Absperrung wie eine Schranke und die Information ob sich ein „Personenhindernis“ davor oder dahinter befindet). Zugleich kann KI genutzt werden, um Bilder zu segmentieren (z. B. Gleisbereich vs. Bereich neben dem Gleis).

Adressierte KI-Arten: Üblicherweise werden heute faltende neuronale Netze (CNN) mit einer gewissen Anzahl an Zwischenschichten (Deep Learning) für die Objekterkennung in Bildern verwendet, weil sie gute Erkennungsraten in kurzer Zeit ermöglichen. Zugleich dürfte eine Sicherheitsargumentation bei dieser Art von KI jedoch vergleichsweise schwierig sein. Einen knappen Überblick über die gängigsten Algorithmen zur KI-basierten Objekterkennung bietet [15]:

- Regionale faltungsneuronale Netze (R-CNN)
- Schnelles R-CNN
- Schnelleres R-CNN
- Regionales vollständig gefaltetes Netzwerk (R-FCN)
- Histogramm der orientierten Verläufe (HOG)
- YOLO (You Only Look Once)
- Einzelschuss-Detektor (SSD)
- Räumliches Pyramiden-Pooling (SPP-net)

Bezüge zu anderen Domänen: Im Straßenverkehr gibt es eine parallele Entwicklung zum hochautomatisierten Fahren, wo neben Hindernissen beispielsweise auch Verkehrsschilder oder Straße/Fahrspuren aus Bildern mit Hilfe von KI erkannt werden. Daneben wird Objekt-/Bilderkennung in sehr vielen weiteren Anwendungsfällen beispielsweise in Robotik, Industrie (Vermessung/Sortierung/Fehlererkennung), Behörden (Gesichtserkennung) oder Alltag (Fotoklassifizierung) eingesetzt.

Beispiel 2: Zustandserfassung und Fehlerdiagnose für Gleise und Weichen per Multisensor-System

Hintergrund

Um die sichere Betriebsführung zu gewährleisten, müssen Eisenbahninfrastrukturunternehmen (EIU) unter anderem den Schienenoberbau (Gleise, Weichen, Schwellen, etc.) überwachen und instandhalten. Die DB Netz AG als Betreiber des öffentlichen Bahnnetzes in Deutschland schreibt in ihrem Regelwerk (DB-Richtlinie 821 [4]) daher Gleisbegehungen in regelmäßigen Intervallen vor; analoge Vorschriften finden sich in den jeweiligen Regularien privater Netzbetreiber. Die vielzähligen einzelnen Überprüfungen finden dabei teils manuell, teils automatisiert statt [5]. Wenn Schäden oder Abweichungen von definierten Schwellwerten (z. B. Gleishohllage, Abstandsmaße) festgestellt werden, müssen entsprechende Maßnahmen ergriffen werden.

Solche *präventiven* Gleisbegehungen, -befahrungen, Weicheninspektionen etc. sind kostenintensiv, da nicht nur Personal und Geräte bezahlt, sondern teils auch die Strecke gesperrt werden muss. Zugleich sind sie nicht besonders effizient, da der gesamte Oberbau vor Ort inspiziert wird, um wenige problematische Stellen zu finden, und da der beste Zeitpunkt für eine Inspektion/Wartung sich selten mit dem durch das Inspektionsintervall gegebenen Zeitpunkt deckt. Die daneben praktizierte *korrektive* Instandhaltung beim Auftreten von Fehlern (z. B. Weichenstörung) ist ebenfalls oft keine effiziente Strategie, weil Fehler häufig das Ergebnis einer graduellen Zustandsverschlechterung sind.

Auf diesem Hintergrund gab es in den letzten Jahren zunehmend Forschung, sowie Bestrebungen der EIU, für ihre Assets inkl. des Oberbaus eine quasi-kontinuierliche Zustandsüberwachung durch relativ kostengünstige Sensorik zu erreichen. Die erfassten Daten (z. B. Messung und Auswertung der Stromkurve von Weichenmotoren [6]) sollen genutzt werden, um zu einer *zustandsbasierten* bzw. im Idealfall sogar *prädiktiven* Instandhaltung zu gelangen. Die Datenauswertung geschieht dazu teils mit Hilfe von verschiedenen KI-basierten Ansätzen [7, 8], die z. B. zur Anomalieerkennung und Fehlerdiagnose eingesetzt werden. Erhofft wird sich durch die Analyse der Sensordaten auch eine verbesserte (Fern)Diagnose bei auftretenden Störungen, eine gezieltere Analyse und Behebung im Feld, sowie eine Reduktion von Fehlalarmen gegenüber heutigen Überwachungsmethoden.

Definition

Beispiel 2 behandelt die Überwachung der Geometrie von Schienen, Weichen und Kreuzungen sowie der Funktionsfähigkeit der Weichen in einem definierten Bereich eines Eisenbahnnetzes. Diese erfolgt indirekt über die Auswertung von Daten verschiedener geeigneter Sensorik, insbesondere:

- Gyrosensorik, die radnah auf einem Regelzug/mehreren Regelzügen angebracht ist, der/die täglich den gesamten Netzbereich befahren. Es werden 3D-Beschleunigungs- und Lagedaten erfasst und per Satellitenortung und Odometrie mit Positionsdaten sowie der aktuellen Zuggeschwindigkeit versehen.
- Stellstromkurven aller Weichenmotoren werden aufgezeichnet.[2]

[2] Ein alternativer Ansatz, der hier nicht weiter betrachtet wird, ist die Weichenüberwachung durch auf Schwellen montierte Vibrationssensorik [9].

Um die Daten normalisieren und besser interpretieren zu können, werden weitere relevante Einflussgrößen erfasst, insbesondere:

- Lufttemperatur, Luftfeuchtigkeit und Niederschläge im Netzbereich;
- Gleis-/Weichen/Kreuzungstemperatur;
- Anzahl der Achsen, die über ein Gleis, eine Weiche oder Kreuzung rollen (z. B. per existierendem Achszähler oder zusätzlicher Erschütterungssensorik am Gleis).

Auf die gewonnenen und ggf. normalisierten Daten werden KI-Verfahren zu mehreren Zwecken angewandt:

- Anomaliedetektion
- Separation der zugseitig aufgenommenen Beschleunigungsdaten nach Rad und Schiene [10]
- Fehlerdiagnose

Ziel ist es, jegliche Inspektionen zur Geometrie- und Weichenfunktionsüberprüfung, für die Personen vor Ort bzw. auf dem Wartungsfahrzeug nötig sind, durch die Überwachung zu ersetzen, Wartungen nur noch zustandsbasiert (und ggf. nach einem Maximalintervall) durchzuführen, sowie bei Unregelmäßigkeiten oder Störungen eine Diagnose zu erstellen bzw. die Suche nach den Ursachen zu unterstützen.

Beispiel 2 umfasst folgende Inspektionsaufgaben:

- Spurweite, Lagefehler, Überhöhung, Verwindung, etc.
- Unebenheiten des Schienenkopfes/Rades, Gleis-/Radrauheit
- Position/Maße Weichenherzstück
- Leichtgängigkeit/Blockierung der Weichenzungen (z. B. Gleitstuhlschmierung, Fremdkörper)
- Funktionsfähigkeit Weichenverschluss
- Elektrischer Anschluss Weichenmotor

Das Beispiel lässt offen,

- wie groß der betrachtet Netzbereich ist,
- ob es sich um ein öffentliches oder privates Netz handelt,
- welche Abtastraten für Beschleunigungs- und Stromkurven die Sensoren liefern (übliche Raten reichen offensichtlich für KI-Anwendungen aus), und
- welche Genauigkeit die Sensordaten besitzen.

Nicht unter Beispiel 2 fallen:

- Andere Oberbaubestandteile wie Schwellen, Schotter, Gleisklammern etc.
- Materialzustand von Gleisen oder Weichenbauteilen.

Charakteristika

Sicherheitsrelevanz: Dass die Schienengeometrie innerhalb definierter Grenzwerte gehalten wird ist neben Abnutzungsaspekten und Reisekomfort vor allem wegen der Entgleisungsgefahr wichtig. In der Praxis treten Entgleisungen aus Gründen mangelnder Wartung der Schieneninfrastruktur in Deutschland kaum auf; neben korrektiver Wartung oder Gleiserneuerung werden auch

Geschwindigkeitsbegrenzungen zum Erhalt der Entgleisungssicherheit eingesetzt. Die Funktion einer Weiche (Umlaufen) hat dagegen kaum sicherheitliche Bedeutung, da über nicht korrekt in Endlage verschlossene Weichen im Stellwerk keine Fahrstraße eingestellt werden kann; die Folgen für den Betriebsablauf können je nach Position der Weiche im Netz jedoch schwerwiegend sein. Dennoch wurde die Weichenfunktion, die hauptsächlich über die Stellstromkurve diagnostiziert wird, hier mit aufgenommen, da verschiedene Abhängigkeiten zwischen Weichengeometrie und -funktion denkbar sind.

Sicherheitsbetrachtung: Das Sicherheitsmanagement der EIU umfasst üblicherweise klare Vorgaben, innerhalb welcher Soll-Wertebereiche Gleisgeometrien akzeptabel sind. Da diese Vorgaben in der Regel bewährt sind, erscheint es sinnvoll, sich weiterhin darauf zu stützen und Abweichungen des Ist-Zustands vom Soll-Zustand der dort verwendeten Parameter per KI zu ermitteln. Trotzdem stellt dies wegen der Neuheit des KI-Einsatzes mutmaßlich eine signifikante Änderung im Eisenbahnsystem dar, die gemäß EU-Recht einer Risikoanalyse bedarf. In diesem Fall könnte die Sicherheit nachgewiesen werden, wenn die Abweichungen mindestens genauso umfassend und zuverlässig wie durch bisher übliche Begehungen/Befahrungen detektiert würden, z. B. durch einen Parallelbetrieb für eine gewisse Zeit und in einem gewissen Umfang. Da eine Diagnose per KI – d. h. das Mapping der Sensordaten auf die Gleisparameter – noch nicht ausgereift zu sein scheint, wäre ein erstes Ziel zu zeigen, dass eine KI-basierte Anomalieerkennung alle zu überwachenden Arten von Abweichungen ebenso zuverlässig erkennt wie bisherige Begehungen/Befahrungen (die genaue Art der Abweichung würde dann gezielt, ggf. auch durch bisherige Methoden, diagnostiziert).

Gründe für KI-Nutzung: Es liegt keine klare Spezifikation vor, wie aus Beschleunigungsdaten oder Stellstromkurven Zustandsinformationen über Gleisgeometrie bzw. Weichenfunktion gewonnen werden können. Dies liegt daran, dass es sich um eine indirekte Erfassung des Zustands handelt, die zudem vielfältige mögliche Schadensbilder und Einflüsse aggregiert. Anomalien lassen sich zwar teils auch direkt anhand von Signalparametern (z. B. Maximalstrom/-beschleunigung, Frequenz von Schwingungen, Dauer zwischen charakteristischen Datenpunkten) ermitteln, KI hat sich jedoch hierfür sowie zur Separation von rad- und schienenbezogenen Effekten aus Beschleunigungsdaten als effizient erwiesen. Bei der Weichendiagnose kann KI dem Menschen helfen, trotz der hohen Komplexität (Anzahl möglicher Probleme, deren Überlappung und gegenseitige Einflüsse) eine zielgerichtete Eingrenzung des Problems zu erreichen. Gegen die Nutzung von KI-Verfahren, die Trainingsdaten benötigen, spricht, insbesondere bei der Weichendiagnose anhand einer Stellstrom-kurve, aktuell ein Mangel an geeigneten solchen Daten, sowohl quantitativ als auch qualitativ.

Adressierte KI-Arten: Für die Anomaliedetektion kann mit unüberwachtem maschinellem Lernen gearbeitet werden, ebenso wie für die Separation von Signalen nach Rad und Schiene [10]. Die Fehlerdiagnose für Weichen kann nach heutigem Stand mit Bayes’schen Netzen sinnvoll unterstützt werden [11]; wenn ausreichend Trainingsdaten vorliegen, ist künftig auch überwachtes maschinelles Lernen für Diagnose und Prognose denkbar.

Bezüge zu anderen Domänen: Anomalieerkennung wird in unterschiedlichsten Domänen angewendet, z. B. um Bankbetrug, strukturelle Defekte, medizinische Probleme oder Fehler in Texten zu erkennen; dabei können sowohl überwachte als auch unüberwachte Lernverfahren verwendet werden. Bayes’sche Netze haben als probabilistische Expertensysteme ebenfalls ein breites Einsatz-spektrum, unter anderem in Bioinformatik, Musteranalyse, Medizin und Ingenieurswissenschaften.

Beispiel 3: Bestimmung von Zugposition und -vollständigkeit per Fiber Optic Sensing

Hintergrund

Heutige Bahnen verwenden als Sicherungsverfahren gegen eine Kollision mit Folgefahrten das Fahren im Blockabstand. Erst wenn der vorausfahrende Zug einen Blockabschnitt vollständig verlassen hat, wird es möglich, ihn als Teil einer Fahrstraße für die Folgefahrt zu reservieren und das Signal, das den Block gegen Befahren schützt, auf Fahrt zu stellen. Zugposition und Zugvollständigkeit werden dabei in gewissen Abständen – in Deutschland meist durch Achszählpunkte – detektiert (Gleisbesetzt- bzw. Gleisfreimeldung); beim Fahren mit dem Zugsicherungssystem ETCS wird die Zugposition mit Hilfe von im Gleis verlegten Balisen ermittelt.

Ein Nachteil des Verfahrens ist die notwendige teure Infrastruktur (Signale und Achszähler sowie ggf. Balisen für jeden Abschnitt); ein anderer die je nach Blockgröße mehr oder weniger starke Einschränkung der Streckenkapazität gegenüber dem Fahren im (absoluten) Bremswegabstand, das eine dichtere Zugfolge ermöglicht. Letzteres Verfahren wird auch „Moving Block“ genannt. Voraussetzung dafür ist jedoch eine (quasi-)kontinuierliche Bestimmung der Zugposition und Zugvollständigkeit. Technologien hierfür werden aktuell, vor allem im Rahmen des europäischen Shift2Rail-Forschungsprogramms, konzipiert; dabei wird auf zugseitige Satellitenortung und auf Spitze und Ende des Zugs verteilte Systeme zur Zugvollständigkeitsdetektion gesetzt.

Eine streckenseitige Alternativlösung für beide Zwecke – Bestimmung von Zugposition und -vollständigkeit – könnte die Fiber Optic Sensing Technologie sein, die seit 2012 für verschiedenste Anwendungszwecke von der DB erforscht wird [12]. Dabei wird regelmäßig ein Lichtsignal durch eine am Gleis befindliche Glasfaser geschickt, deren Länge im zweistelligen Kilometerbereich liegt. Akustische Ereignisse / Vibrationen am Gleis bewirken eine Veränderung des Signals, das anschließend aufgezeichnet und ausgewertet wird. Dabei kann auch auf den Ort des Ereignisses entlang der Glasfaser geschlossen werden, je nach Sampling-Rate mit Genauigkeiten von unter einem Meter. Aktuelle Forschung [13] zeigt, dass die Drehgestelle bzw. Achs- oder Drehgestellcluster eines Zuges im Signal aufgelöst werden können, was die Kontrolle ihrer Anzahl und damit der Zugvollständigkeit erlaubt. Daneben lässt sich gemäß [13] die Geschwindigkeit des Zuges je nach Verfahren mit Genauigkeiten von weniger als 5 oder sogar 1km/h bestimmen.

Definition

Beispiel 3 umfasst die regelmäßige Bestimmung (z. B. alle 2 Sekunden) der Zugpositionen und Zugvollständigkeiten aller auf einem zweigleisigen Streckenabschnitt von 20km Länge verkehrenden Zügen per Fiber Optic Sensing. Die zeit- und ortsbezogenen Rohdaten werden dabei vorprozessiert, z. B. um die Amplituden näherer und entfernterer Signale zu normalisieren und die Signale zu glätten. Schließlich können aus der Intensität des Signals über dem Streckenabschnitt zum aktuellen Zeitpunkt t die Positionen der Zugmitten und damit die Zugpositionen bestimmt werden (auch ohne KI).

Da die Signalqualität einer einzelnen Intensitätskurve über der Zeit (an einem festen Ort) oft nicht zur Prüfung der Zugvollständigkeit ausreicht, und weil die Kurven für Orte, an denen sich der Zug aktuell befindet, noch nicht von allen Drehgestellen/Achsen überfahren wurden, werden dafür mehrere Kurven von hinter dem Zug liegenden Orten betrachtet. Die Kurven werden an der Ankunftszeit des ersten Drehgestells/Achsclusters aneinander ausgerichtet und eine Durchschnittsbildung

durchgeführt. Für eine ausreichend genaue und einheitliche Bestimmung dieser Ankunftszeit wird eine KI eingesetzt. Schließlich wird die Anzahl der Peaks der resultierenden Kurve detektiert und mit der erwarteten bzw. vorherigen Anzahl der Drehgestelle/Achscluster des Zuges verglichen, um die Zugvollständigkeitsinformation zu erhalten.

Einen Überblick über das System gibt Abbildung 2. Wegen des hohen Datenvolumens soll davon ausgegangen werden, dass der/die Rechner in Nähe des Signalmessungssystems platziert und die Übertragung der Daten kabelgebunden stattfindet. Die Resultate (Zugpositionen, ermittelte Anzahl Achscluster) werden ebenfalls kabelgebunden an das übergeordnete Kontrollsystem (Stellwerk) weitergegeben.

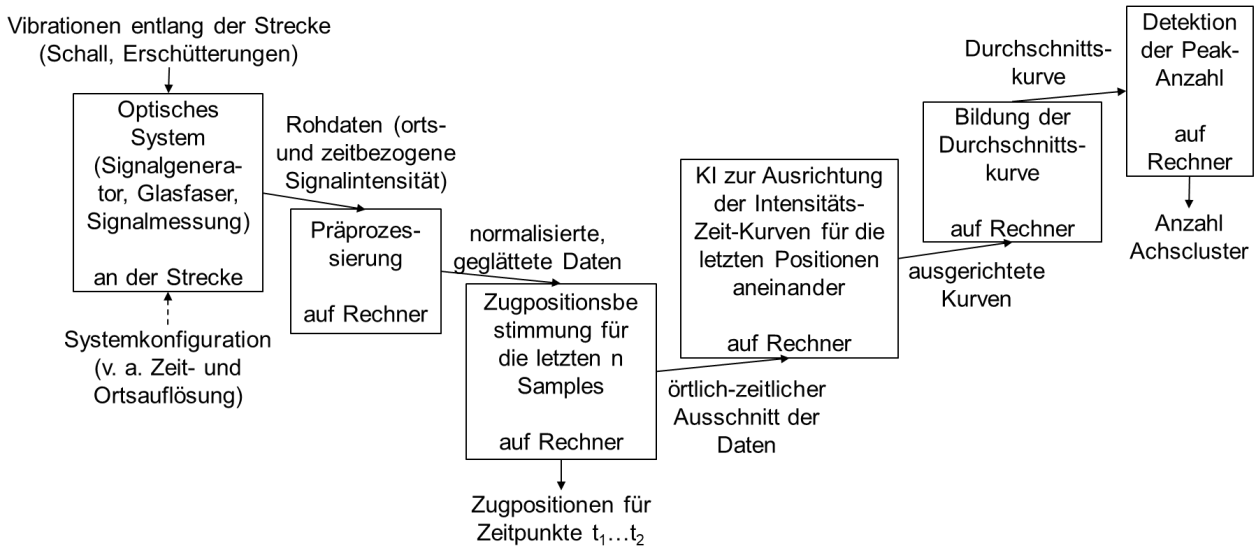


Abbildung 2: Skizze des Systems zur Bestimmung von Zugposition und -vollständigkeit

Ziel ist es, auf dem Streckenabschnitt das Fahren im Moving Block auf Basis der erhaltenen Zugpositionen und Zugvollständigkeitsinformationen zu realisieren. Dabei wird von einem zentralen System ähnlich [17] ausgegangen, das Stellwerks- und ETCS Radio Block Center-Funktionalität integriert und basierend auf dem Status der Strecke Fahrerlaubnisse an Züge in seinem Bereich erteilt, die sich abzüglich eines Sicherheitspuffers bis zum (zuletzt gemeldeten) Ende des vorausfahrenden Zuges erstrecken können. Während die Meldung in [17] ETCS-konform per Positionsreport über Funk vorgesehen ist, wird sie im vorliegenden Fall durch das Fiber Optic Sensing System übernommen. Gleisfreimeldeanlagen sind in Übereinstimmung mit [17] lediglich für besonders kritische Abschnitte wie Weichen(bereiche) bzw. zur Detektion des Eintritts/Austritts eines Zuges in/aus einen/einem Bereich vorgesehen.

Beispiel 3 bezieht sich geografisch auf Deutschland und umfasst Folgendes:

- Bestimmung der Zugposition und der Zugvollständigkeit aus den Fiber Optic Sensing Daten;
- beliebige Züge und Achs-/Drehgestellkonfigurationen auf der betrachteten Strecke.

Damit Züge sicher durch das System erkannt werden können, muss während einer Fortbewegung des Zuges das charakteristische, sich fortbewegende Signalmuster der rollenden Achscluster ohne längere Aussetzer erkennbar sein. D. h. insbesondere muss sichergestellt werden, dass (a) das charakteristische Signalmuster in ausreichender Stärke erzeugt wird, (b) Störsignale (inkl. Zügen auf

Parallelgleisen) beschränkt werden und (c) ein ähnliches Signalmuster nicht von einer anderen, sich entlang der Schienen fortbewegenden Quelle (Straßen- oder landwirtschaftliche Fahrzeuge) erzeugt wird. Eine genauere Definition/Einschränkung der Systemumgebung für eine erste Version eines Systems könnte daher wie folgt aussehen (die angegebenen konkreten Werte basieren auf Bahnwissen, Recherchen und Überlegungen des Autors; es wird keine Gewähr für die Eignung für reale Systementwicklungen übernommen):

- Bahninfrastruktur
    - Maximal zweigleisige Strecke mit kleineren/mittleren Bahnhöfen (<=4 parallele Gleise)
    - Oberbau ist als feste Fahrbahn ausgeführt
    - Glasfaser liegt (ohne dass Zugspannung auf sie wirkt) weitgehend parallel zum Gleis (ohne größere Schwankungen im Verhältnis von Gleislänge und entsprechender Kabellänge) in einem abgedeckten Kabelschacht im Boden (geschützt vor direkter mechanischer Einwirkung) nahe den Gleisen (i. d. R. <=8 m Abstand zur Gleismitte des weiter entfernteren Gleises)
- Zug
    - Achscluster weisen einen maximalen Abstand d1 > Spurweite/4 zwischen den Achsen auf
    - Achscluster weisen einen minimalen Abstand d2 > 2*d1 sowie einen maximalen Abstand d3 mit d2 < d3 < 25 m zwischen ihren sich am nächsten liegenden Achsen auf
    - Maximale Zuglänge 400 m
    - Maximalgeschwindigkeit 300 km/h
- Betrieb
    - Minimaler Abstand aufeinanderfolgender Züge 500 m
    - Keine längeren (länger als 200 m) Schleichfahrten <10 km/h
    - Bei Überholung durch einen Zug auf dem Nachbargleis wird die Position des sicheren Zugendes des überholten Zuges ab Beginn des Überholvorgangs festgehalten bis beide Züge sie vollständig passiert haben (bedeutsam für Folgefahrten des überholten Zuges, die entsprechend Abstand halten müssen)
    - Bei Überholung eines Zuges auf dem Nachbargleis wird die Position des sicheren Zugendes des überholenden Zuges ab dem Moment, in dem beide Zugenden gleichauf sind, durch die des Zugendes des überholten Zuges ersetzt, bis der Überholvorgang abgeschlossen ist (bedeutsam für Folgefahrten des überholenden Zuges, die entsprechend Abstand halten müssen)
- Erschütterung, Vibrationen, Umgebungsschall
    - Keine ausgedehnten (länger als 200 m) Wartungsarbeiten an der Strecke
    - Keine ausgedehnten (länger als 200 m) Verkehrsinfrastrukturen oder maschinenbewirtschafteten landwirtschaftlichen Flächen entlang der Strecke im Abstand von weniger als 25 m zur Strecke
    - Keine Nutzung in Erdbebengebieten
    - Keine bekannten dauerhaften Quellen von starkem Lärm und Erschütterungen entlang der Schienen
- Witterungsbedingungen
    - Umgebungstemperaturbereich von -20°C bis +45°C

Weitere Umgebungsbedingungen können je nach Positionierung des Systems – neben der Glasfaser befinden sich voraussichtlich auch Signalmessung (Wandler optisch -> elektronisch) und -verarbeitung (Rechner) mehr oder weniger nah an der Strecke – relevant sein. Ihre Beachtung ist jedoch standardmäßig für Komponenten der Leit- und Sicherungstechnik geboten, weswegen von einer genaueren Definition abgesehen wird. Hier zählen Staub/Schmutz/Feuchtigkeit entsprechend des dauerhaften Einsatzes in Außenbereichen und elektromagnetische Bedingungen für elektronische Geräte [18]. Auch adaptierte Betrachtungen zu Störeinflüssen zwischen Zug und streckenseitiger Sensorik nach dem Vorbild von [20] scheinen sinnvoll.

Das Beispiel lässt offen (Festlegung technischer Details nur bei Bedarf und Konsens der Projektpartner),

- welche Art von Fiber Optic Sensing zum Einsatz kommt (z. B. einfach vs. true phase),
- welche genaue zeitliche und räumliche Auflösung die Rohdaten haben,
- welche Rohdatenmenge dabei erzeugt wird,
- welche Qualität die Sensordaten haben (z. B. Signal-zu-Rauschen-Verhältnis), und
- wieviel Zeit die Vorprozessierung der Daten benötigt.

Nicht unter Beispiel 3 fallen:

- Die Betrachtung des Einflusses verschiedener baulicher Gegebenheiten (Details der Oberbauweise, der Glasfaserführung, etc.) auf das Signal;
- die Bestimmung der Zuggeschwindigkeit aus den Sensordaten;
- die Ermittlung weiterer Informationen aus den Sensordaten.

Um möglichst vollständige Anforderungen an das System zur Zugpositionierung und Zugvollständigkeitsdetektion formulieren zu können, erscheint es wie bereits für Beispiel 1 aufgrund der nicht exakt spezifizierbaren Systemeingabe sinnvoll, allgemein relevante Eingabeparameter zu identifizieren und diese später in KI-Trainings- und Testdaten in einer gewissen Granularität abzudecken. Im Unterschied zu Beispiel 1 ist allerdings die Korrelation zwischen den Parametern der Realität und denen der gemessenen Daten (Kamerabild in Beispiel 1 bzw. zeitlich-räumliches Signal in Beispiel 3) weniger direkt; die Entfernung des Zuges von der Messeinrichtung, Achsabstände, Zuggeschwindigkeit, konkrete Quellen von Störsignalen wie Züge auf einem Parallelgleis etc. übersetzen sich in Signalstärke, Frequenz, Schwingungsmuster, Rauschen, Störsignal etc. Beide Ebenen haben ihre Vor- und Nachteile bei der Nutzung für die Formulierung von Anforderungen, und es ist wahrscheinlich, dass zur Erzeugung von Trainings- und Testdaten auch beide Ebenen genutzt werden (Kombination von eingemessenen und synthetisch erzeugten/kombinierten Signalen). Um am Ende jedoch dem Einsatz im realen Umfeld gerecht zu werden, scheint es zunächst einmal wichtig, auf der Ebene der Realität Anforderungen (an das Gesamtsystem inklusive KI) zu spezifizieren. Wichtige Parameter inklusive möglicher Grenzwerte finden sich in der obigen beispielhaften Umgebungsdefinition; die Granularität der Testwerte der meist kontinuierlichen Größen könnte durch eine in der Realität erwartete Verteilung sowie Randwerte definiert werden. Hinzu kommen Kalibrierungsparameter des Fiber Optic Sensing Systems. Für letztere können Abhängigkeiten von anderen Parametern oder ein Kalibrierungsverfahren spezifiziert werden, um das System in verschiedenen Strecken- und Betriebskontexten nutzen zu können.

Auch bzgl. der Ausgabe unterscheidet sich Beispiel 3 von Beispiel 1: statt einem Vektor von Objektklassen mit Konfidenzen werden hier Skalarwerte durch die KI berechnet, und ob diese die „gewünschte“ (nicht näher definierte) Feinausrichtung der Samples aneinander ergeben, kann erst nach der Berechnung der Achsclusterzahl durch den Vergleich mit der tatsächlichen Achsclusterzahl des Zuges ermittelt werden. Verlässlich testen lässt sich die KI also erst als Teil des Gesamtsystems. Trotzdem lässt sich die Ausgabe der KI bereits anhand verschiedener Maße beurteilen (z. B. Vergleiche von aneinander ausgerichteten Samples auf minimale Differenzen ihrer Peakmitten oder auf maximalen Flächenüberlapp, oder Bewertung der Glätte der Mittelwertkurve mehrerer ausgerichteter Samples).

Überlegungen zu Systemarchitektur und zuverlässigkeitssteigernden Maßnahmen

Im Gegensatz zu den Beispielen 1 und 2 hat die KI im Beispiel 3 eine viel stärker eingegrenzte Aufgabe, nämlich die präzise einheitliche Berechnung des Ankunftszeitpunktes eines Zuges an einem festen Ort aus der Kurve der Signalintensität über der Zeit an diesem Ort. Da diese Berechnung für viele Kurven durchgeführt wird, aus denen anschließend eine Durchschnittskurve erzeugt wird (vgl. Abbildung 2), hat das Verfahren bereits eine gewisse Robustheit gegenüber Fehlern einzelner KI-Anwendungen. Darüber hinaus fließen die KI-Ergebnisse lediglich in die Zugvollständigkeitsbetrachtung bzw. die detektierte Anzahl von Achsclustern ein; ein allgemein plausibler Wertebereich hierfür beschränkt sich auf ganze Zahlen zwischen eins und wenigen hundert; und in jedem spezifischen Fall einer Detektion für einen Zug ist die Zahl der erwarteten Achscluster sogar exakt bekannt.

Entlang der Verarbeitungskette bieten sich diverse weitere Möglichkeiten der Absicherung bzw. Plausibilisierung:

- Eine Zugtrennung geht nicht nur mit einer reduzierten Zahl an Achsclustern des vorderen Zugteils einher, sondern bewirkt in der Regel auch, dass ein zweiter Zugteil entsteht. Ab einem gewissen Abstand zum ersten Zugteil würde dieser bei der Zugpositionsbestimmung (vgl. Abbildung 2) als eigener Zug und somit unabhängig von der KI detektiert.
- Die Fortbewegung eines Zuges lässt sich für kleine Zeitdifferenzen recht genau vorhersagen. Weicht die von der KI berechnete Ankunftszeit an einem Ort von einem entsprechenden Zeitfenster ab, kann eine fehlerhafte KI-Berechnung identifiziert werden.
- Das durch die KI gelöste Problem ist prinzipiell auch durch einen klassischen Peak-Finder-Algorithmus lösbar [13]. Dieser besitzt zwar eine längere Laufzeit, eine (verzögerte) Online-Verifikation der KI-Ergebnisse ist auf diese Weise aber möglich.
- Als weiterer komplementärer Ansatz kann Anomaliedetektion (mit oder ohne KI) auf die Reihe der Kurven angewendet werden, die die Eingabe für die KI zur Ausrichtung der Kurven darstellen. Auch damit sollte ein Verlust der Zugvollständigkeit detektierbar sein.
- Maßnahmen wie diversitäre KI-Algorithmen und Strukturen neuronaler Netze oder Mehrkanaligkeit (vgl. Ausführungen zu Beispiel 1) sind möglich, werden aber angesichts der zuvor genannten Optionen als (unnötig?) teure Maßnahmen eingeschätzt.

Für die weitere Erhöhung der Robustheit des Verfahrens bieten sich ebenfalls verschiedene Ansätze an:

- Das System sollte Fremdsignale/starkes Rauschen, z. B. bei Zugbegegnungen oder von Baumaschinen, erkennen und entsprechende Messwerte temporär von der Auswertung

ausnehmen bzw. bei dauerhaften Fremdsignalen an der Zugposition melden, dass es nicht funktionsfähig ist.
- Das System könnte statt dem einfachen Vergleich mit der erwarteten Achscluster-Anzahl Änderungen der Anzahl über die Zeit detektieren; eine Zugtrennung könnte erst dann als detektiert gelten, wenn sich die geänderte Anzahl über einige wenige Auswertungszyklen stabilisiert.

Für ein konkretes System bzw. eine Systemkonfiguration wird es wichtig sein, die Einsatzbedingungen präzise zu beschreiben. Dies beinhaltet u.a. Folgendes:

- Es sollte klar definiert sein, welche Achskonstellationen durch das System erkannt werden können (Mindestabstand zwischen unterscheidbaren Achsclustern/Achsen, Mindestabstände von unterscheidbaren Zügen), und wie die Glasfaser zu installieren (Abstand vom Gleis, Bettung, tolerierbare Längendifferenzen) und das System zu konfigurieren ist.
- Es sollte klar definiert sein, unter welchen Bedingungen das System betrieben werden darf (z. B. keine ausgedehnten Quellen von Erschütterungen entlang der Glasfaser etwa durch Großbaustellen oder Erdbeben, Umgang mit Rangiertätigkeiten oder Stärken/Schwächen von Zügen).
- Eine sichere Behandlung von Schleichfahrten/Zugstillstand muss gewährleistet sein. Übernahme/Übergabe auf/in Nachbarsysteme (Anfang und Ende der überwachten Strecke, Abzweigungen) sollte klar beschrieben und der Ablauf verifiziert sein.

Charakteristika

Sicherheitsrelevanz: Im Fall der vorliegenden Anwendungen können insbesondere falsche Positionsmeldungen oder Zugvollständigkeitsmeldungen zu Zugkollisionen und somit katastrophalen Folgen führen. Ausfälle der Meldungen sind zwar nicht sicherheitsrelevant, können jedoch gerade bei Moving Block starke negative Auswirkungen auf den Betriebsablauf haben.

Sicherheitsbetrachtung: Für Kernkomponenten bzw. -funktionen der Sicherungstechnik von Vollbahnen wird bzgl. zufälliger Fehler das Sicherheitsniveau SIL4 gefordert, d.h. eine Fehlerrate von höchstens 10^{-9} pro Stunde; für die hier betrachteten Funktionen üblicherweise sogar 10^{-10}. Dies betrifft die verwendete Hardware wie Glasfaser, Generierung und elektronische Verarbeitung der Lichtimpulse. Softwareanteile müssen zur Vermeidung systematischer Fehler den Vorgaben der EN 50128 [14] für die höchsten Software-SIL-Stufen 3/4 entsprechen, was bei der Verarbeitung sehr komplexer Sensordaten und erst recht bei KI-Nutzung schwer zu argumentieren sein dürfte (vgl. Tabellen in Anhang A der Norm). Andererseits wird die KI im vorliegenden Beispiel lediglich für eine kleinere Teilaufgabe genutzt und es sind diverse Möglichkeiten denkbar, die Datenqualität an verschiedenen Stellen des Verarbeitungsprozesses automatisiert zu überprüfen/sicherzustellen, resultierende Zugpositionen/Zugvollständigkeitsinformationen zu plausibilisieren sowie Ergebnisfehler zu begrenzen. Generell zu bedenken ist, dass ein generischer Sicherheitsnachweis verschiedene Arten der Verlegung der Glasfaser und ein jeweils korrektes Mapping der Glasfaserposition auf die Gleisposition mitbetrachten muss, ebenso wie das geschwindigkeitsabhängige Verhältnis von Vibrationen/Geräuschen, die durch die rollenden Räder einerseits und andere Quellen andererseits verursacht werden.

Gründe für KI-Nutzung: Im Vergleich zu einem klassischen Peak-Finder kann KI Ankunftszeiten der Züge an einem festen Ort der Strecke schneller und flexibler (für verschiedene Zuggeschwindigkeiten) ermitteln sowie in einer Weise, dass die mit Hilfe der Ankunftszeiten gebildete Durchschnittskurve gleichförmiger ist und somit eine bessere Detektion der Drehgestelle/Achscluster erlaubt [13]. Offensichtlich bildet die KI eine besser geeignete Definition von „Ankunftszeit“ ab als die zu einfache/zu starre analytische Definition bei der Peak-Finder-Nutzung.

Adressierte KI-Arten: Für die Bestimmung der Ankunftszeit eignet sich ein neuronales Netz, das mit echten sowie synthetischen Daten trainiert wurde [13].

Bezüge zu anderen Domänen: Fiber Optic Sensing ist eine schon länger am Markt befindliche Technologie, die häufig zur Überwachung ausgedehnter Infrastrukturen (z. B. Pipelines, Grenzverläufe) oder Messung in größeren Gebieten (z. B. seismische Aktivität) verwendet wird. Die Nutzung von KI zur Analyse von Zeitreihen ist in vielzähligen Gebieten zu finden, beispielsweise in den Natur- und Ingenieurswissenschaften oder in Finanzwesen und Wirtschaft.

Literatur

[1] DB Netze: Ril 408 Fahrdienstvorschrift. Gültig ab 15.12.2019. Verfügbar unter https://fahrweg.dbnetze.com/fahrweg-de/kunden/nutzungsbedingungen/regelwerke/betrieblich-technisch_regelwerke/betrieblich_technisches_regelwerk-4613476?contentId=1369926, zuletzt abgerufen am 13.1.2021.

[2] Thales: Künstliche Intelligenz für die autonom fahrende Stadtbahn. News-Artikel der Thales-Website vom 15.2.2019. Verfügbar unter https://www.thalesgroup.com/de/deutschland/news/kuenstliche-intelligenz-fuer-die-autonom-fahrende-stadtbahn, zuletzt abgerufen am 14.1.2021.

[3] IEC 62267 Railway applications - Automated urban guided transport (AUGT) - Safety requirements. Edition 1.0, Ausgabedatum 2009-07.

[4] DB Netze: Ril 821 Oberbau inspizieren. 2016. Internes Regelwerk der DB Netz AG.

[5] P. A. Wilfling: Der Weg zur smarten Weiche. Masterarbeit, TU Graz, Mai 2017. Verfügbar unter https://pure.tugraz.at/ws/portalfiles/portal/12105001/DerWegzurSmartenWeiche.pdf, zuletzt abgerufen am 15.1.2021.

[6] DB: Digitale Weichendiagnose mit DIANA. Artikel der DB-Website, aktualisiert am 17.6.2020. Verfügbar unter https://inside.bahn.de/digitale-weichendiagnose-diana/, zuletzt abgerufen am 14.1.2021.

[7] N. Kornfeld, A. Luber, A. Leich, M. Kaiser, L. A. Schubert, J. C. Groos: Zustandsüberwachung der Bahninfrastruktur mit KI. In: EIK - Eisenbahn Ingenieur Kompendium 2020. DVV Media Group, 2019, Seiten 318-335. Verfügbar unter https://elib.dlr.de/132958/.

[8] B. Baasch, J. Heusel, J. C. Groos, S. Shankar: Eingebettete Zustandsüberwachung der Gleisinfrastruktur: Entwicklung und Erprobung von eingebetteten Multi-Sensor-Systemen für die kontinuierliche Zustandsüberwachung der Gleisinfrastruktur im operativen Betrieb. Der Eisenbahningenieur EI 12/2019, Seiten 6-8. Verfügbar unter https://elib.dlr.de/129363/.

[9] KONUX: Produktbeschreibung auf der Unternehmenswebsite. Verfügbar unter https://www.konux.com/de/loesung/, zuletzt abgerufen am 19.1.2021.

[10] B. Baasch, M. Roth, S. Schulz, J. C. Groos: An unsupervised machine learning approach to extract wheel and track health status indicators from train-borne accelerometer data. ESREL2020 PSAM15, 01. – 05. Nov. 2020, Venedig, Italien. Abstract verfügbar unter https://elib.dlr.de/134005/.

[11] T. Neumann, D. Narezo Guzman, J. C. Groos: Transparente Fehlerdiagnose bei Weichenstörungen mittels Bayes'scher Netze. SIGNAL + DRAHT 12/2019, 111 (12), Seiten 23-31. Verfügbar unter https://elib.dlr.de/128850/.

[12] F.A.Z.: So hört die Bahn Kabeldiebe oder Tiere. Artikel vom 22.10.2017. Verfügbar unter https://www.faz.net/aktuell/technik-motor/technik/fiber-optic-sensing-so-hoert-die-bahn-kabeldiebe-oder-tiere-15247562.html, zuletzt abgerufen am 20.1.2021.

[13] S. Kowarik, M.-T. Hussels, S. Chruscicki, S. Münzenberger, A. Lämmerhirt, P. Pohl, M. Schubert: Fiber Optic Train Monitoring with Distributed Acoustic Sensing: Conventional and Neural Network Data Analysis. Sensors 2020, 20, 450; doi:10.3390/s20020450

[14] DIN EN 50128:2012 Bahnanlagen – Telekommunikationstechnik, Signaltechnik und Datenverarbeitungssysteme – Software für Eisenbahnüberwachungs- und Steuerungssysteme; Deutsche Fassung EN 50128:2011.

[15] A. Zola: Objekterkennung und -Segmentierung. Blog-Artikel auf der Website der clickworker GmbH vom 14.09.2021. Verfügbar unter https://www.clickworker.de/kunden-blog/objekterkennung-und-segmentierung/, zuletzt abgerufen am 3.1.2022.

[16] G. Zauner, R. Slabihoud, M. Bürger, F. Auer: Kamerabasierte Objekterkennung mit tiefen neuronalen Netzwerken zur automatischen Erfassung der Gleisinfrastruktur. ETR 4/2019, Seiten 45-49.

[17] X2Rail-3 WP4: Deliverable D4.2 Moving Block Specifications. Deliverable D4.2 des X2Rail-3 Projekts, 2020. Verfügbar unter https://projects.shift2rail.org/s2r_ip2_n.aspx?p=X2RAIL-3

[18] DIN EN 50121: 2017 Bahnanwendungen - Elektromagnetische Verträglichkeit; Deutsche Fassung EN 50121:2017

[19] DIN EN 50155:2018 Bahnanwendungen - Elektronische Einrichtungen auf Schienenfahrzeugen; Deutsche Fassung EN 50155:2017

[20] DIN EN 50238:2020 Bahnanwendungen - Kompatibilität zwischen Fahrzeugen und Gleisfreimeldesystemen; Deutsche Fassung EN 50238-1:2019

Sicherheitsnachweisführung für KI-Verfahren

AP5 – Wechselwirkung zwischen KI und Sicherheitsnachweisführung

Systementwicklung nach EN 50126

4) Festlegung der Systemanforderungen

Funktion

- Die Funktionstüchtigkeit einer KI kann über eine suffiziente Anzahl von Szenarien abgefragt werden
- Diese sollten eine Vielzahl von erwartbaren Ereignissen und die Anforderungen an die KI zu großen Teilen abdecken
- Weiterhin sollten Extremfälle vorbereitet werden, die trotzdem von der KI erkannt werden müssen.
- Gleichermaßen wird ein Teil der verfügbaren Daten als Testdatensatz zurückgehalten
- Die Performance auf den, dem Netz unbekannten, Testdaten ist in Indikator für das Verhalten in einer realen Umgebung

Robustheit

- Die Robustheit von KI kann mittels externer Bibliotheken getestet werden
- Dabei wird die Anfälligkeit des Netzes gegenüber kleinen Änderungen überprüft

Wartbarkeit

- Es wurde beschlossen das die KI in der Produktion nicht weiter trainiert wird.
- Dadurch fallen für die KI selbst keine Anforderungen zur Wartung an.
- Sollte der Bedarf entstehen die KI mit neuen Daten zu trainieren, muss diese entsprechend neu oder weiter trainiert werden.
    - Durch die nicht-durchschaubaren Änderungen während des Trainings müssen trotzdem alle Anforderungen erneut geprüft und abgenommen werden.

Leistung

- Die Leistung der KI wird üblicherweise über unterschiedliche Metriken überprüft
    - Bahnspez. Bsp. 1: Objekt-Detection
        - Average Precision (AP)
        - mean Average Precision (mAP) – Die AP wird für jede Klasse individuell berechnet, mAP ist das Mittel über alle Klassen
        - AP50, AP75, etc.
        - Precision x Recall Kurve
    - Bahnspez. Bsp. 3: Regression

        - Übliche Metriken sind z.B. der MSE (Mean Squared Error) oder MAE (Mean Absolute Error)
- Die Metriken werden während des Trainings auf ein kleineres Validierungsset angewendet und sind ein guter Indikator für die Leistung des neuronalen Netzwerkes
- Entscheidend sind hier allerdings die Metriken, der KI auf den zurückgehaltenen Testdaten

Effizient

- Effizienz ist keine Kenngröße im Bereich KI
- Hier könnte aber die Parameteranzahl (Komplexität), sowie Berechnungszeit der KI angeführt werden
- Mehr Parameter → mehr Berechnungen → längere Berechnungszeit
- Schafft die KI einen vollständigen Vorwärtsdurchlauf bis zur Deadline im Echtzeitbetriebssystem?
- Bahnspez. Bsp. 3: Wie viel FPS können verarbeitet werden? Was ist die Responsetime der KI?
- Bahnspez. Bsp. 1: Kann eine kleinere KI ähnliche Ergebnisse liefern? Ist eine tiefere KI genauer?

Sicherheit

- Die Aspekte der Sicherheit und Robustheit überschneiden sich (zumindest von meiner Begriffsdefinition)
- Hierzu zählt die Anfälligkeit der KI gegenüber von gezielten kleinen Änderungen, mit denen versucht wird die Ausgabe zu beeinflussen (Adversarial Examples)
- Es existieren externe Bibliotheken z.B. CleverHans, welche die Effektivität bekannter Evasion-Techniken auf ein Netzwerk automatisiert Testen können
- Quantifiziert kann die Sicherheit hier evtl. aus der Differenz der entsprechenden Metriken zwischen natürlicher Eingabe und angegriffener Eingabe werden.

Schnittstellen

- Die in den Beispielen in Betracht gezogene KI benötigt in der Produktion von allem zwei Schnittstellen
    - Eine Eingabe (die erste Schicht des DNN), welche die Daten im vorher definierten Format entgegen nimmt
        - Bahnspez. Bsp. 1: Je nach Algorithmus ist wird ein vorverarbeitets Bild direkt als Weite x Höhe x Farbkanäle Matrix an das Netzwerk übergeben
        - Bahnspez. Bsp. 3: Das im Paper beschriebene Netz bekommt als Eingabe 3000 Zeitsamples in denen sich der Zug an einer Position x befindet

- Eine Ausgabe (die letzte Schicht des DNN), welche das Ergebnis der mathematischen Berechnungen die auf den Input angewendet werden ist
  - Bahnsepz. Bsp. 1: Je nach Algorithmus wird üblicherweise eine Vektor zurückgegeben, welcher sich über alle möglichen Klassen erstreckt
    - Jeder Index de s Vektors korrespondiert dabei statisch mit einer Klasse (wird bereits zum Training festgelegt)
    - Im Index befindet sich ein Confidence-Wert der angibt wie zuversichtlich das Netz ist, dass die Eingabe zur entsprechenden Klasse gehört
  - Bahnspez. Bsp. 3: Die Ausgabe des beschriebenen Netzwerkes ist ein Skalar (die Zeitverschiebung um alle Achsen, auf die gleiche Zeit auszurichten)

Betriebliche Anwendung

- Üblicherweise laufen KI-Applikationen auf einer graphischen Recheneinheit (GPU)
- Diese unterscheiden sich allerdings je nach Modell in den Faktoren Speichergröße, Berechnungen pro Sekunde etc.
- Es muss also eine GPU gewählt werden, die …
  - das Netzwerk vollständig in den Speicher laden kann
  - genügend FLOPS hat um die Berechnung im Bezug auf die Deadline durchzuführen

Zu berücksichtigende Umgebungsbedingungen

- **-**

6) Entwurf und Implementierung

- Unter Entwurf fällt hier ggf. noch die Auswahl der neuronalen Netzwerkes
- Supervised Trainingsvorgang eines arbiträren DNN
  - Aufteilen des vorhandenen Datensatzes in Trainings- und Testset
    - Je nach Anzahl der Daten mit 80/20, 90/10, etc. Splittung
    - Bei kleineren Datensätzen sollte auf Fairness überprüft werden, bzw. ob die Samples der individuellen Klassen in einem angemessenen Maße auf Trainings- und Testset verteilt sind um später aussagekräftige Metriken im Bezug auf das Testset zu erhalten
    - Ggf. können mittels Datenaugmentation weitere Samples erstellt werden um die Daten künstlich zu erweitern
  - Trainingsphase
    - Ggf. werden die Eingaben durch verschiedene Vorverarbeitungsstuffen angepasst.
    - Während des Trainings werden die Samples des Trainingssets in sog. Batches unterteilt. Die Größe der Batch hat einen direkten Einfluss auf das Lernverhalten der

KI, da der Fehler über mehr Daten hinweg berechnet wird, was optimalerweise bei die Verallgemeinerungsfähigkeit der KI unterstützt.

- Vor dem Training werden die Gewichte des Netz entweder nach einem bestimmten Algorithmus oder zufällig festgelegt.
- Bevor das eigentliche Training startet muss eine Fehlerfunktion (Loss) definiert werden. Die Funktion schätzt die Qualität der KI ein (Bsp. MSE).
- Während des Trainings soll der Fehler des Netzwerkes auf den Trainingsdaten minimiert werden.
- Das Training selbst ist daher ein Optimierungsproblem, zur Lösung des Problems gibt es eine Reihe von Algorithmen, bekannt sind z.B. der Stochastic Gradient Descent (SGD) oder Adam.
    - Diese sog. Optimizer verfügen oft über eine Reihe von unterschiedlichen Parametern (z.B. Momentum), welche aber abhängig vom jeweiligen Algorithmus sind.
- Die eigentliche Trainingsschleife besteht aus zwei Schritten:
    - Forward-Pass: Die Eingaben (Batch) wird vom Netz verarbeitet und eine entsprechende Ausgabe berechnet.
    - Backpropagation: Anhand des berechneten Fehlers wird zurück durch das Netz propagiert, dabei werden die Gewichte des NN basierend auf ihrem Fehler angepasst. Die Lernrate entscheidet dabei wie groß die Änderung innerhalb einer Iteration ist.
- Mit jeder Iteration wird also versucht die Gewichte so zu verschieben, das der Fehler minimal ist.
- Die Schleife wird dann so lange wiederholt bis entweder eine bestimmte Zahl oder eine Abbruch-Kondition (z.B. genügend Validierungsgenauigkeit) erreicht wurde.

9) Systemvalidierung

- Analytischer Nachweis des gelernten schwierig. Es existieren Techniken die zeigen auf was genau die in der Eingabe achtet.
- Techniken wie z.B. LIME oder SHAP, können zeigen welcher Teil der Eingabe die Entscheidung des Netzwerkes in die eine oder andere Richtung beeinflusst haben.
- Nachweis ist evtl. ein Punkt für ein weitere Forschung
- Zum Nachweis der Funktionalität können Szenarios vorbereitet werden. Diese sollten die Funktionsanforderungen an die KI abdecken (natürlich keine vollständige Abdeckung möglich)
- Das Testen von Szenarien kann über Simulation durchgeführt werden.
- Wenn die definierten Szenarien erfolgreich abgeschlossen werden können, sollte ein praktischer Test z.B. auf einer Teststrecke oder einer beliebigen Bahn.

- Im praktischen Test können die Ergebnisse der KI über einen längeren Zeitraum überwacht und überprüft werden.
- Die Unabhängigkeit zwischen Lerndaten und Testdaten wird in erster Linie dadurch erreicht, dass nur die Lerndaten in die Trainingsphase integriert werden. Die Testdaten bleiben damit für das Netz unbekannt und können zur Validierung herangezogen werden.
- Bei sehr kleinen Datensätzen besteht hier die Möglichkeit über Kreuzvalidierung eine bessere Aussage über die Leistung der KI zu erhalten.
- Wenn es Klassen mit wenigen Daten geben sollte, kann bei der Trennung der Lern- und Testdaten sichergestellt werden, dass diese gleichmäßig verteilt sind.
- Äquivalenzklassen könnten die Komplexität der KI verringern, da dadurch Parameter in der letzten Schicht entfallen. Was für Auswirkungen dies auf die Performance der KI ist experimentell festzustellen.

11) Betrieb, Instandhaltung und Leistungsüberwachung

- Es ist ein System notwendig, welches in der Lage ist die Eingabedaten bis zur festgelegten Deadline zu berechnen und an die nächste Komponente zu übergeben.
- Mein Vorschlag für Instandhaltung wäre:
  - Während des Betriebs werden weiter Daten gesammelt und eventuelle neue Situationen werden an der aktuellen KI getestet.
  - Wenn die KI diese ebenfalls mit hinreichender Genauigkeit bewältigt ist keine „Instandhaltung“ notwendig, ansonsten muss evtl. weiteres Training durchgeführt werden
- Basierend auf dem trainierten Modell können Analysen zu potentiellen Fehlern durchgeführt werden. Dafür wird das Modell auf Daten angewendet. Die falsch erkannten Samples werden dann genauer analysiert.
  - Welche Klassen werden prozentual eher falsch vorhergesagt? Wie könnten die falschen Samples zusammenhängen?
- Weiterhin können vorausgewählte Samples geprüft werden, diese sollten zum einen Teil für Menschen einfach erkennbar sein und dementsprechend auch vom Modell richtig erkannt werden.
- Durch eine Robustheitsanalyse (z.B. CleverHans) können bereits Fehlerquellen erkannt werden.
- Durch Techniken wie LIME (Iteratives Abdecken von Bereichen der Eingabe) kann außerdem getestet werden was genau für Features für das Netz wichtig sind, so können ebenfalls frühzeitig Fehler erkannt werden.
- Weiterhin können die Aktivierungen der Schichten eines Netzwerkes überprüft werden. Schichten die oftmals einen Gradienten von 0 haben könnten schlecht trainiert sein.
- Um schlechte Verhältnisse zu simulieren könnte etwa Noise in die Eingabe eingeführt werden.

- Eine weitere Möglichkeit Erkennung von potentiellen Fehlern könnte über zusätzliche Software geschehen, so dass z.B. Ergebnisse mit geringer Confidence als Fehler erkannt werden und vom System entsprechend behandelt werden.

1 Systementwicklung nach EN 50126

1.1 Vorbemerkung

Für die Entwicklung einer sicherheitsrelevanten KI-Anwendung gilt der Systementwicklungsprozess nach EN 50126 und EN 50129. Für die Softwareentwicklung gelten zusätzlich die Anforderungen an den Entwicklungsprozess gemäß EN 50128.

Keine dieser Normen berücksichtigt den Einsatz von KI. Aus diesem Grund wird nachfolgend die Entwicklung einer KI-Anwendung auf den normativen Entwicklungsprozess abgebildet. Dabei wird dargestellt, wie die wesentlichen normativ geforderten Inhalte der einzelnen Phasen auf eine KI-Anwendung angewendet werden können und welche Fragestellungen bzw. offenen Punkte gegenwärtig noch nicht hinreichend beantwortet sind.

1.2 Phase Systemkonzept

In der Phase Systemkonzept werden die Strategie und Ziele der Systementwicklung festgelegt. Es werden der Anwendungsbereich, der Kontext und der Zweck des Systems definiert. Ein wesentlicher Punkt ist die Analyse und Beschreibung der künftigen Systemumgebung, d.h. der technischen, physikalischen, geografischen und klimatischen Umgebungsbedingungen, der betrieblichen Randbedingungen, der Schnittstellen zu vorhandenen Systemen sowie der zu beachtenden Gesetze, Normen und Regelwerken.

In dieser Phase erfolgt auch eine Analyse und Gegenüberstellung ähnlicher oder vergleichbarer Systeme. Dabei werden die Anforderungen an diese Systeme sowie deren technische und betriebliche Parameter sowie die Sicherheitsziele dieser Systeme ermittelt und auf ihre Anwendbarkeit bzw. Übertragbarkeit auf das neu zu entwickelnde System untersucht.

Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen. Im AP4 wurden beispielhaft Systemkonzepte bahnspezifischer Anwendungen skizziert.

1.3 Systemdefinition und betrieblicher Kontext

Hier erfolgt eine Beschreibung des künftigen Systems einschließlich dessen Funktionalität und einer ersten Architekturskizze, aus der die Komponenten des Systems und die Verteilung der Aufgaben auf diese Komponenten hervorgehen.

Es wird beschrieben, wie das System im Betrieb angewendet werden soll. Hierzu gehören Betriebsverfahren, Betriebsarten, Use Cases zu verschiedenen betrieblichen Szenarien und die Instandhaltungsstrategie.

Außerdem werden die klimatischen, mechanischen und elektrischen Einsatzbedingungen sowie die betrieblichen Umgebungsbedingungen spezifiziert.

Es erfolgt eine Beschreibung der Systemgrenzen sowie der Schnittstellen und Wechselwirkungen zu anderen technischen Systemen und zum Menschen, d.h. dem Bediener, Instandhalter und Nutzer des Systems.

Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen.

1.4 Risikoanalyse und -beurteilung

In dieser Phase werden unerwünschte Ereignisse während des Betriebs des Systems und die daraus resultierenden Gefährdungen ermittelt. Hierzu werden zunächst alle Fehler identifiziert, die Ursache derartiger Ereignisse sein könnten und die Fehlerfolgen hinsichtlich ihrer Kritikalität bewertet. Zu jeder Fehlerfolge wird das akzeptable Risiko ermittelt. Ist das tatsächliche Risiko höher als der akzeptierte Wert, werden Maßnahmen zur Risikoreduktion festgelegt, um sicherzustellen, dass das Risiko auf das akzeptable Maß reduziert wird.

Das akzeptierte Risiko wird als SIL (safety integrity level) und THR (tolerable hazard rate) angegeben. Der ermittelte SIL bestimmt den Umfang der Maßnahmen zur Vermeidung von systematischen Fehlern im Entwicklungsprozess (z.B. Dokumentation, Analysen, Tests, Verifikation und Validierung). Die THR legt die zulässige Rate gefährlicher Ausfälle fest und bestimmt somit die technische Gestaltung des Systems (z.B. Bauelementeauswahl, Redundanzprinzipien, Selbsttests, Überwachung im Betrieb).

Die Vorgehensweise zur Ermittlung des akzeptierten Risikos wird in der CSM-Verordnung (EU) 402/2013 und (EU) 2015/1136 vorgeschrieben. Gängige Verfahren sind die Übernahme vorhandener Werte durch die Anwendung anerkannter Regeln der Technik oder der Vergleich mit einem bereits existierenden und akzeptierten Referenzsystem. Wenn beides nicht möglich ist, muss das akzeptierte Risiko explizit ermittelt werden (z.B. über die Verfahren Risikograph oder Risikomatrix).

Für technische Funktionen der Eisenbahnsicherungstechnik gilt die Vornorm DIN VDE V 0831-103 als anerkannte Regel der Technik. Dort werden die THR für typische Funktionen hergeleitet.

Für die Beispielanwendungen aus AP4 können die Anforderungen wie folgt hergeleitet werden:

<u>Hinderniserkennung auf/neben dem Gleis per Kamera</u>

- Vergleich mit der Funktion „Gefahrenraum freimelden“
- Die zu betrachtende Gefährdung ist: „Aufprall auf ein Hindernis im Gefahrenraum“.
- Die Bewertung der Barrieren erfolgt konservativ mit der Punktzahl 1.
- THR = 1E-06 h^{-1} (entspricht SIL1)

Die TU Berlin verfolgt ein Forschungsprojekt „Risikoakzeptanzkriterien für den automatisierten Fahrbetrieb (ATO-RISK)“. Dazu zählt auch die Ermittlung des tolerablen Risikos und die Gewährleistung der mindestens gleichen Sicherheit, wenn man vom Fahren mit Triebfahrzeugführer auf automatischen Betrieb (ATO - Automatic Train Operation) wechselt. Nach Abschluss dieses Projekts können ggf. die dort ermittelten Werte übernommen werden.

<u>Bestimmung von Zugposition und -vollständigkeit per Fiber Optic Sensing</u>

Wird die KI hier nur zur Gleisfreimeldung verwendet und gibt es über der Gleisfreimeldung ein zusätzliches Sicherungssystem, welches die Ergebnisse plausibilisiert (z.B. Streckenblock), dann gilt:

- Vergleich mit der Funktion „Gleisabschnitt auf Freisein überwachen (Zugstraße)“
- Die zu betrachtende Gefährdung ist: „Zusammenstoß mit anderen Fahrzeugen“.
- Die Bewertung der Barrieren erfolgt mit der Punktzahl 2.
- THR = 1E-08 h^{-1} (entspricht SIL3)

Soll jedoch die vollständige Sicherung durch KI realisiert werden, dann gilt:

- Vergleich mit der Funktion „Schutz gegen Gegenfahrten sicherstellen (Zugfahrstraße)“
- Die zu betrachtende Gefährdung ist: „Zusammenstoß mit anderen Fahrzeugen“.
- Es gibt keine Barrieren.
- THR = 1E-09 h^{-1} (entspricht SIL4)

Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen. Mit der Festlegung eines SIL > 0 ergeben sich jedoch Anforderungen an die „sichere Softwareentwicklung“ der KI-Anwendung. Die Unterschiede zur Entwicklung einer „klassischen“ Software werden in den nachfolgenden Phasen betrachtet.

1.5 Festlegung der Systemanforderungen

In dieser Phase werden die funktionalen und technischen Anforderungen an das zu entwickelnde System vollständig definiert. Hierzu gehören z.B.:

a) Funktionen
b) Robustheit und Wartbarkeit
c) Leistung und Effizienz
d) Sicherheit
e) Schnittstellen
f) betriebliche Anwendung
g) zu berücksichtigende Umgebungsbedingungen (Klima, Mechanik, elektrische Bedingungen, EMV, physische und IT-Security Zugriffe)

Die Anforderungen müssen eindeutig, vollständig, widerspruchsfrei, korrekt, identifizierbar und prüfbar sein.

Im Gegensatz zur klassischen Entwicklung mit präzisen, vollständigen Anforderungen soll KI dort eingesetzt werden, wo die Anforderungen nicht vollständig formuliert werden können (oder nur mit unverhältnismäßig hohem Aufwand). Die Anforderungsspezifikation der KI wird daher eher eine Beschreibung der zu beherrschenden Szenarien und dem jeweils erwarteten Ergebnis sein.

Offener Punkt für Nachfolgeprojekte

Gegenwärtig gibt es keine Kriterien, nach denen die Anforderungsspezifikation einer KI-Anwendung erstellt werden kann, so dass diese die Kriterien Vollständigkeit und Prüfbarkeit erfüllt.

1.6 Architektur und Aufteilung der Systemanforderungen

In dieser Phase wird eine Systemarchitektur entwickelt, die in der Lage ist, die spezifizierten Anforderungen zu erfüllen. Die Anforderungen werden den Teilsystemen und Komponenten zugewiesen und es werden die Schnittstellen zwischen diesen Teilsystemen und Komponenten spezifiziert.

Die Architektur der in AP4 beschriebenen Beispielapplikationen besteht im Wesentlichen aus:

- Sensoren
- einer Rechnerplattform inklusive Betriebssystem
- der generischen KI-Software
- der spezifischen KI-Software (d.h. den erlernten Daten und Parametern der generischen KI-Software)
- der Ausgabeschnittstelle an die übergeordnete Anwendung

Die Systemhardware inklusive der implementierten Redundanz- und Überwachungsmechanismen muss in der Lage sein, die geforderte THR zu erfüllen. Die Software (Betriebssystem, generische und spezifische KI-Software) muss hinsichtlich enthaltener systematischer Fehler den geforderten SIL erfüllen.

Hierzu sind bereits in der Architektur die Prinzipien zum Erreichen der geforderten Sicherheit festzulegen (z.B. fail-safe Hardware, Redundanz, Hard- und Software-Diversität).

Für COTS-Komponenten (z.B. Betriebssysteme oder Standardsoftware) werden in den Normen Kriterien definiert, nach denen deren Betriebsbewährung und Eignung für einen entsprechenden SIL beurteilt werden kann.

Offener Punkt für Nachfolgeprojekte

Die als generische KI-Software infrage kommenden Systeme sind gegenwärtig noch proprietäre Einzelanwendungen, die nicht die Kriterien der Betriebsbewährung und Eignung für COTS-Komponenten erfüllen. Um diese Software einzusetzen, ist eine umfassende Validierung erforderlich, in der nachgewiesen wird, dass die KI-Software frei von systematischen Fehlern ist oder die Wahrscheinlichkeit des Erkennens eines konkreten Sachverhalts begrenzt. Es gibt gegenwärtig kein Verfahren, mit dem eine hinreichende Validierung mit vertretbarem Aufwand möglich wäre.

Der Lernprozess zur Entwicklung der spezifischen KI-Software entspricht nicht den Anforderungen an die Softwareentwicklung gemäß EN 50128. Eine Möglichkeit zur Verringerung der Anforderungen an diesen Prozess wäre der Einsatz diversitärer KI-Software (z.B. Lernen mit diversitären Daten) und Vergleich der Ergebnisse oder eine zusätzliche Plausibilisierung der Ergebnisse der KI. Dies wurde jedoch in den vorangegangenen Arbeitspaketen nicht weiter untersucht.

1.7 Entwurf und Implementierung

In dieser Phase erfolgt der Entwurf der Teilsysteme und Komponenten, so dass sie die an sie gestellten Anforderungen erfüllen.

Die Hardwareentwicklung wird aus den weiteren Betrachtungen ausgeklammert, da sie sich nicht von der Entwicklung konventioneller sicherer Systeme unterscheidet.

Der Entwurf und die Implementierung der KI-Applikation entsprechen der „Lernphase“. Am Ende der Entwicklung muss diese Lernphase abgeschlossen sein, d.h. der Zustand der KI-Software wird „eingefroren“ und ein weiteres Lernen muss technisch ausgeschlossen werden.

Offener Punkt für Nachfolgeprojekte

Die klassische Softwareentwicklung nach EN 50128 wird bei einer KI-Anwendung durch den Lernprozess ersetzt. Folgende Fragen konnten in den vorangegangenen Arbeitspaketen noch nicht vollständig geklärt werden:

1) Wie kann die Aufgabenstellung an die KI-Anwendung auf einen vollständigen Satz prüfbarer Anforderungen und Lerndaten abgebildet werden?

2) Nach welchen Kriterien kann das Ende des Lernprozesses festgelegt werden, bzw. welche Anforderungen werden an Datenmenge, Datenqualität und Datendiversität gestellt, um sicher zu sein, dass die Aufgabenstellung am Ende des Lernprozesses vollständig und korrekt beherrscht wird?

1.8 Herstellung

In dieser Phase erfolgt die Fertigung der Komponenten und Teilsysteme. Da es hier keine Unterschiede zu einer konventionellen Entwicklung gibt, wird diese Phase nicht weiter betrachtet.

1.9 Integration

In dieser Phase erfolgt die Integration der Teilsysteme und Komponenten zum Gesamtsystem. Es wird nachgewiesen, dass das integrierte System korrekt zusammenwirkt, um die vorgesehene Funktion zu erfüllen.

Da es hier zunächst keine Unterschiede zu einer konventionellen Entwicklung gibt, wird diese Phase nicht weiter betrachtet.

Maßnahmen zur Integration von KI-Software werden jedoch relevant, wenn zum Erreichen der geforderten Sicherheit eine unabhängige Plausibilisierung der KI-Funktion angewendet wird oder Rückfallebenen beim Erkennen von Fehlern der KI-Software vorgesehen wurden.

1.10 Systemvalidierung

Während der Systemvalidierung erfolgt die Prüfung und Bestätigung, dass das betrachtete System für den vorgesehenen Verwendungszweck geeignet ist und die festgelegten Anforderungen erfüllt. Die wesentlichen Mittel der Validierung sind Analysen und Tests.

Im Rahmen der Systemvalidierung sind zwei Sachverhalte nachzuweisen:

1) Hat die spezifische KI-Software die Anforderungen (Lerninhalte) korrekt im Sinne der Aufgabenstellung interpretiert? Hierzu ist ein analytischer Nachweis erforderlich.

2) Werden alle Anforderungen an die KI-Software vollständig und korrekt erfüllt? Dieser Nachweis erfolgt durch entsprechende Tests.

<u>Offene Punkte für Nachfolgeprojekte</u>

Die folgenden Fragen konnten in den vorangegangenen Arbeitspaketen nicht geklärt werden:

1) Wie und in welchem Umfang kann analytisch nachgewiesen werden, „was“ die KI-Software gelernt hat?

2) Nach welchen Kriterien kann ein hinreichender Testumfang definiert werden, um eine vollständige Anforderungsüberdeckung durch Tests zu erreichen?

3) In welcher Testumgebung können bzw. müssen die Tests ausgeführt werden (Labor, Simulation, Feld)?

4) Wie kann eine hinreichende Unabhängigkeit zwischen „Lerndaten“ und „Testdaten“ erreicht werden, um systematische Fehler aufgrund gleicher und ggf. unzureichender Datenquellen zu vermeiden?

1.11 Systemabnahme

In dieser Phase erfolgt die abschließende Begutachtung bzw. Bewertung des Systems und dessen Abnahme durch den Betreiber. Grundlage hierfür bilden die Nachweise der vorangegangenen Phasen und ggf. ergänzende Prüfungen oder Tests.

Die Anforderungen an diese Phase gelten gleichermaßen für konventionelle Systeme und KI-Anwendungen.

1.12 Betrieb, Instandhaltung und Leistungsüberwachung

Zu dieser Phase gehören:

a) Betrieb des Systems
b) Schulung des Personals
c) Instandhaltung des Systems
d) Fehlermanagement (Analyse, Bewertung, Korrektur)
e) Führen des Gefährdungslogbuchs

Offene Punkte für Nachfolgeprojekte

Es müssen entsprechende Verfahren (Anleitungen) für Betrieb und Instandhaltung der generischen und spezifischen KI-Software entwickelt werden. Hierzu sind folgende Fragen zu klären:

1) Wodurch unterscheiden sich die Verfahren zum Betrieb und der Instandhaltung eines KI-Systems von denen eines konventionellen Systems?

2) Wie kann der Fehlermanagementprozess eines KI-Systems definiert werden (d.h. das Verfahren zum Erkennen von Fehlern der KI-Software, zur Analyse der Fehlerursache und der Fehlerauswirkungen)

3) Welche Maßnahmen sind zur Fehlerbehebung der KI-Software erforderlich (z.B. weiteres Lernen oder „Löschen“ oder „Korrigieren“ fehlerhaften Wissens)?

1.13 Außerbetriebnahme

In der Phase Außerbetriebnahme erfolgen das Abschalten des Systems am Ende seiner Nutzung und das Entfernen aus seiner Systemumgebung.

Die Anforderungen an diese Phase gelten zunächst gleichermaßen für konventionelle Systeme und KI-Anwendungen. Die KI-Software ist in ihren Ausgangszustand vor Beginn der Lernphase zu versetzen.

2 Sicherheitsnachweis nach EN 50129

2.1 Vorbemerkung

Im Sicherheitsnachweis wird dargelegt und begründet, wie das System die gestellten Anforderungen an Funktionalität und Sicherheit erfüllt. Aufbau und Struktur des Sicherheitsnachweises sind in der Norm EN 50129 festgelegt. Für Softwareanwendungen sind zusätzlich die Anforderungen der EN 50128 zu berücksichtigen.

Der „technische“ Nachweis erfolgt im Teil „Technischer Sicherheitsbericht“. Nachfolgend wird skizziert, wie die wesentlichen normativ geforderten Inhalte an den Technischen Sicherheitsbericht auf eine KI-Anwendung angewendet werden können und welche Fragestellungen bzw. offenen Punkte aus Sicht der Nachweisführung gegenwärtig noch nicht hinreichend beantwortet sind.

2.2 Abschnitt Einleitung

In diesem Abschnitt erfolgt ein Überblick über das System und den Systementwurf und eine Darstellung der Prinzipien, auf denen sich die Sicherheit des Systems abstützt. Die Inhalte leiten sich aus den Ergebnissen der Phasen „Systemkonzept“ bis „Festlegung der Systemanforderungen“ ab.

Offene Punkte für Nachfolgeprojekte

Die Voraussetzung für die Sicherheitsnachweisführung ist neben einer hinreichenden System-beschreibung eine Klärung der auf KI anwendbaren Sicherheitsprinzipien.

In den vorangegangenen Arbeitspaketen wurden keine speziellen Sicherheitsprinzipien für KI-Software identifiziert. Nach Abschluss der Lernphase befindet sich die KI-Anwendung in einem „eingefrorenen“ Zustand. Ihr Verhalten wird nicht von stochastischen Größen beeinflusst, so dass sie zu jedem Zeitpunkt mit identischen Eingangsdaten auch identische Ergebnisse liefert. Damit unterscheidet sie nichts von einer herkömmlichen Software-Anwendung.

Der Unterschied zur klassisch entwickelten Software besteht darin, dass der Lernprozess nicht mit den Methoden der klassischen Softwareentwicklung verifiziert werden kann.

Es konnte noch nicht geklärt werden, wie die Qualität des Lernprozesses quantifiziert werden kann.

Ein möglicher Lösungsweg, auch mit nicht vollständig verifizierter Software eine hinreichende Sicherheit zu erzielen, ist der Einsatz diversitärer Software. Hierzu müssten diversitäre KI-Anwendungen mit unterschiedlichen Lernansätzen eingesetzt oder diese mit unterschiedlichen Arten von Lerndaten trainiert werden.

Es konnte noch nicht geklärt werden, inwieweit dies möglich ist und in welchem Umfang bei dieser Lösung Common-Cause Effekte berücksichtigt werden müssen.

2.3 Abschnitt Nachweis des korrekten funktionalen Verhaltens

Der wesentliche Inhalt dieses Abschnitts ist der Nachweis zur Erfüllung der funktionalen Anteile der System-Anforderungsspezifikation (d.h. der funktionalen betrieblichen Anforderungen) und der Sicherheits-Anforderungsspezifikation (d.h. der funktionalen Sicherheitsanforderungen).

Hierzu gehören der Nachweis der korrekten Hardware-Funktionalität und der Nachweis der korrekten Software-Funktionalität. Letzterer ist für die KI-Anwendung relevant. Die anzuwendende Norm ist die EN 50128.

Der Nachweis der korrekten Softwarefunktionalität erfolgt durch Analysen und Tests.

Angewendet auf eine KI-Anwendung bedeutet dies, dass in einem analytischen Nachweis zu zeigen ist, dass die KI-Software „das Richtige“ gelernt hat. Die Ergebnisse dieses analytischen Nachweises sind durch Tests zu plausibilisieren. An die Tests werden folgende Anforderungen gestellt:

- vollständige Anforderungsüberdeckung
- Unabhängigkeit der Testentwicklung von der Systementwicklung

Offene Punkte für Nachfolgeprojekte

Der analytische Nachweis erfolgt bei herkömmlicher Software über eine durchgehende Anforderungsverfolgung von der Spezifikation über das Design bis zum Code. Dieser Weg ist bei einer KI-Anwendung nicht möglich. Hierzu muss z.B. geklärt werden, wie aus den Parameterbelegungen der KI-Software am Ende der Lernphase auf die erlernten Inhalte geschlossen werden kann.

Eine weitere offene Frage ist, wie auf Basis einer typischerweise „unscharfen“ Anforderungsspezifikation der KI-Anwendung ein Nachweis der vollständigen Anforderungsüberdeckung geführt werden kann.

Um die Unabhängigkeit der Testentwicklung von der Systementwicklung zu erreichen, ist es notwendig, die Unabhängigkeit des Lernprozesses und der Lerndaten vom Testprozess mit Testumgebung und Testdaten nachzuweisen.

Alle verbleibenden Unschärfen oder Fehler der KI-Software sind systematische Fehler. Sie beeinflussen direkt den erreichbaren SIL. In der Norm EN 50129 wird gefordert, dass Fehlzustände aufgrund systematischer Fehler erkannt und zu einer sicherheitsgerichteten Ausfallreaktion führen müssen. Dies kann z.B. durch eine Plausibilisierung der Ergebnisse durch eine zusätzliche Überwachungseinrichtung erfolgen. Dies wird nicht in allen Anwendungsfällen möglich sein und schränkt die Einsetzbarkeit von KI bei hohen Sicherheitsanforderungen (SIL) entsprechend ein.

Die generische KI-Software wird im Sicherheitsnachweis als „pre-existing Software“ behandelt. Hierzu muss es gewisse Qualitätsanforderungen erfüllen. Gegenwärtig gibt es keine professionelle KI-Software, mit denen ein solcher Nachweis durchführbar ist.

2.4 Abschnitt Ausfallauswirkungen

In diesem Abschnitt ist insbesondere die Ungefährlichkeit von Hardwareausfällen nachzuweisen. Für KI-Applikationen ist der Teil „Schutz vor systematischen Fehlern“ relevant. Wenn davon ausgegangen werden muss, dass die KI-Software aufgrund der Unmöglichkeit des vollständigen Nachweises des korrekten funktionalen Verhaltens systematische Fehler enthält, dann ist hier zu zeigen, wie solche systematischen Fehler erkannt und in einen sicheren Zustand überführt werden. Eine Möglichkeit hierzu ist eine zusätzliche Überwachungseinrichtung, welche Fehler in den Ergebnissen der KI erkennt und daraufhin eine sichere Reaktion einleitet.

Wird zum Schutz vor gleichgerichteten systematischen Fehlern in der KI-Software der Ansatz diversitärer KI-Software verfolgt, dann ist hier die Unabhängigkeit der diversitären KI zu zeigen. Dies betrifft insbesondere den Schutz vor gleichgerichteten systematischen Fehlern. Dieser kann z.B. erreicht werden durch eine Kombination von:

- diversitärer generischer KI-Software
- diversitären Lernprozessen
- diversitären Lerndaten

Offene Punkte für Nachfolgeprojekte

Es ist gegenwärtig nicht geklärt wie eine hinreichende Diversität (Unabhängigkeit) der verschiedenen verfügbaren generischen KI-Software nachgewiesen werden kann. Gleiches gilt für die Entwicklung diversitärer Lernprozesse und Lerndaten.

2.5 Abschnitt Betrieb mit externen Einflüssen

In diesem Abschnitt erfolgt der Nachweis, dass das entwickelte System auch unter Einwirkung der anzunehmenden externen Einflüsse seine betrieblichen Anforderungen und seine Sicherheitsanforderungen erfüllt. Externe Einflüsse sind z.B. Umgebungsbedingungen wie Temperatur, Luftfeuchte, mechanische und elektrische Einflüsse. Je nach Art der verwendeten Sensoren kommen ggf. auch optische oder akustische Einflüsse hinzu.

Der Nachweis, dass die KI-Software unabhängig von den Umgebungsbedingungen korrekte Ergebnisse liefert, erfolgt überwiegend durch Tests.

Offene Punkte für Nachfolgeprojekte

Ein erhoffter Vorteil der Anwendung von KI ist das Identifizieren charakteristischer Merkmale der Eingangsgrößen trotz nicht vollständig beschreibbarer Umgebungsbedingungen. Auch hier stellt sich daher die Frage, wie ein hinreichender Lern- und Testdatensatz definiert werden kann, um eine vollständige Überdeckung der zu erwartenden Umgebungsbedingungen zu erreichen. Dieser Testdatensatz muss wieder ebenfalls „unabhängig“ von den Lerndaten entwickelt werden.

2.6 Abschnitt Sicherheitsbezogene Anwendungsbedingungen

In diesem Abschnitt erfolgt die Definition aller Regeln, Bedingungen und Einschränkungen, die bei der Anwendung des Systems zu beachten sind, damit dessen Sicherheit gewährleistet ist. Insbesondere müssen hier die aus den Restriktionen der KI-Anwendung resultierenden Einschränkungen und Anwendungsregeln hergeleitet und definiert werden.

Offene Punkte für Nachfolgeprojekte

Es ist anzunehmen, dass KI gegenwärtig noch nicht in der Lage ist, die geforderten Funktions- und Sicherheitsziele vollständig zu erreichen.

Bezogen auf die Anwendungsbeispiele der vorangegangenen Arbeitspakete konnte noch nicht geklärt werden, welche Restriktionen der KI hier zu erwarten sind und durch welche Maßnahmen diese kompensiert werden können.

2.7 Abschnitt Ergebnisse der Sicherheitserprobung

Im Anschluss an den theoretischen Nachweis durch Analysen und Tests erfolgt die Sicherheitserprobung. Diese dient dazu, das Vertrauen in das neu entwickelte System zu stärken. Ihr Ziel ist eine hinreichende Erprobung des Systems unter allen relevanten Betriebsbedingungen.

Zu beachten ist, dass die Sicherheitserprobung nicht mit den Tests zum Nachweis des korrekten funktionalen Verhaltens und des Betriebs mit externen Einflüssen gleichzusetzen ist. Diese Tests müssen vor Beginn der Erprobung erfolgreich abgeschlossen sein. Es genügt daher nicht, die KI im Feld zu „erproben“ und darauf dann die Sicherheit zu begründen.

Man unterscheidet eine Sicherheitserprobung mit und ohne Sicherheitsverantwortung des zu erprobenden Systems. Bezogen auf die KI-Anwendung ist es daher möglich, das System parallel zum bereits existierenden System zu erproben und während dieser Phase die Ergebnisse des KI-Systems mit denen des bestehenden Systems zu vergleichen.

Offene Punkte für Nachfolgeprojekte

Es wurde noch nicht untersucht, über welchen Zeitraum und unter welchen Bedingungen eine Sicherheitserprobung durchgeführt werden muss, damit am Ende der Erprobung ein hinreichendes Vertrauen in das System vorliegt. Dies steht insbesondere im Zusammenhang mit der noch ungeklärten Frage einer hinreichenden Testabdeckung während der Systemvalidierung.

3 Resümee

Die Untersuchungen im Rahmen des Forschungsprojekts KI-bezogene Test- und Zulassungsmethoden haben gezeigt, dass es möglich ist, die Entwicklung einer KI-Entwicklung auf den Entwicklungsprozess gemäß EN 50126, EN 50128 und EN 50129 abzubilden.

Es sind jedoch noch Fragen bzw. offene Punkte vorhanden, die im Rahmen von Nachfolgeprojekten zu klären sind. Diese betreffen insbesondere:

- Wie kann eine Anforderungsspezifikation für KI-Anwendungen so erstellt werden, dass diese das Kriterium „Vollständigkeit“ erfüllt?
- Wie kann ein analytischer Nachweis der Anforderungserfüllung erfolgen, d.h. die Frage beantwortet werden „WAS“ die KI gelernt hat und ob damit alle Anforderungen erfüllt werden?
- Wie kann ein hinreichender Satz an Lern- und Testdaten zum vollständigen Nachweis der Anforderungsüberdeckung erstellt werden?
- Wie kann die Unabhängigkeit von Lern- und Testdaten gewährleistet und nachgewiesen werden?

Mögliche Lösungsansätze zur Beherrschung sind die Entwicklung diversitärer KI-Anwendungen und die Überwachung der KI-Anwendung durch eine zusätzliche Überwachungseinrichtung.

Für den Einsatz zweier diversitärer KI-Anwendungen sind folgende Fragen bzw. offene Punkte zu klären:

- Wie kann die Unabhängigkeit der verwendeten generischen KI-Software nachgewiesen werden?
- Wie kann die Unabhängigkeit der Lernprozesse beider KI-Anwendungen nachgewiesen werden?
- Wie kann die Unabhängigkeit der Lerndaten beider KI-Anwendungen nachgewiesen werden?
- Verbleiben trotz des diversitären Ansatzes noch Common-Cause Effekte und wie können diese identifiziert werden?