Schlussbericht

Siemens Mobility GmbH

Sichere KI am Beispiel fahrerloser Regionalzug

Förderkennzeichen: 19I21039B

SIEMENS

Siemens Mobility GmbH

Krauss-Maffei-Straße 2

80997 München

Deutschland

Finanziert von der Europäischen Union
NextGenerationEU

Gefördert durch:
Bundesministerium für Wirtschaft und Energie
aufgrund eines Beschlusses des Deutschen Bundestages

SIEMENS

Schlussbericht

Safe.trAIn

Sichere KI am Beispiel fahrerloser Regionalzug

Vorhabensbezeichnung:	safe.trAIn – Sichere KI am Beispiel fahrerloser Regionalzug
Zuwendungsempfänger:	Siemens Mobility GmbH
Förderkennzeichen:	19I21039B
Laufzeit des Vorhabens	01.01.2022 bis 31.03.2025
Autoren/Ansprechpartner:	Meike Meller Dr. Thomas Waschulzik Dr. Kristian Weiß

Das diesem Schlussbericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministerium für Wirtschaft und Energie (BMWE) unter dem Förderkennzeichen 19I21039B gefördert.
Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Berichterstatter. Die Form des Berichtes entspricht den Nebenbestimmungen für Zuwendungen auf Kostenbasis des Bundesministers für Wirtschaft und Energie an Unternehmen der gewerblichen Wirtschaft für Forschungs- und Entwicklungsvorhaben (NKBF98).
Wir danken unseren Konsortialpartnern und insbesondere dem Projektträger für die vertrauensvolle Zusammenarbeit.

Abbildungsverzeichnis:

Tabellenverzeichnis:

1 Kurzdarstellung

1.1 Aufgabenstellung

Die Digitalisierung und Automatisierung des Zugbetriebes stellt einen wesentlichen Hebel für die Erreichung der europäischen Klimaschutzziele dar. Um die Verkehrswende zu erreichen, ist die Erhöhung der Attraktivität des Schienenverkehrs ein wesentlicher Faktor. Der Einsatz fahrerloser Schienenfahrzeuge bietet hierfür eine Vielzahl von Vorteilen, die sowohl ökonomische, ökologische als auch soziale Aspekte umfassen:

- Entgegenwirken dem Fachkräftemangel bei Triebfahrzeugführer:innen (Tf)
- Steigerung der Effizienz durch optimierte Betriebsabläufe
- Steigerung der Attraktivität durch höhere Taktzeiten
- Steigerung der Flexibilität bei der Fahrplan-Gestaltung
- Reduktion von Emissionen durch energie-optimiertes Fahren
- Steigerung der Verfügbarkeit

Basierend auf der IEC 62290-1 werden die Stufen der Automatisierung von Zügen in GoA (Grade of Automation)-Level definiert.

- GoA0 – Kein Automatisierungsgrad
- GoA1 – Manuelle Steuerung mit Zugbeeinflussung
- GoA2 – Teilautomatisierter Betrieb
- GoA3 – Fahrerloser Betrieb
- GoA4 – Vollautomatisierter Betrieb

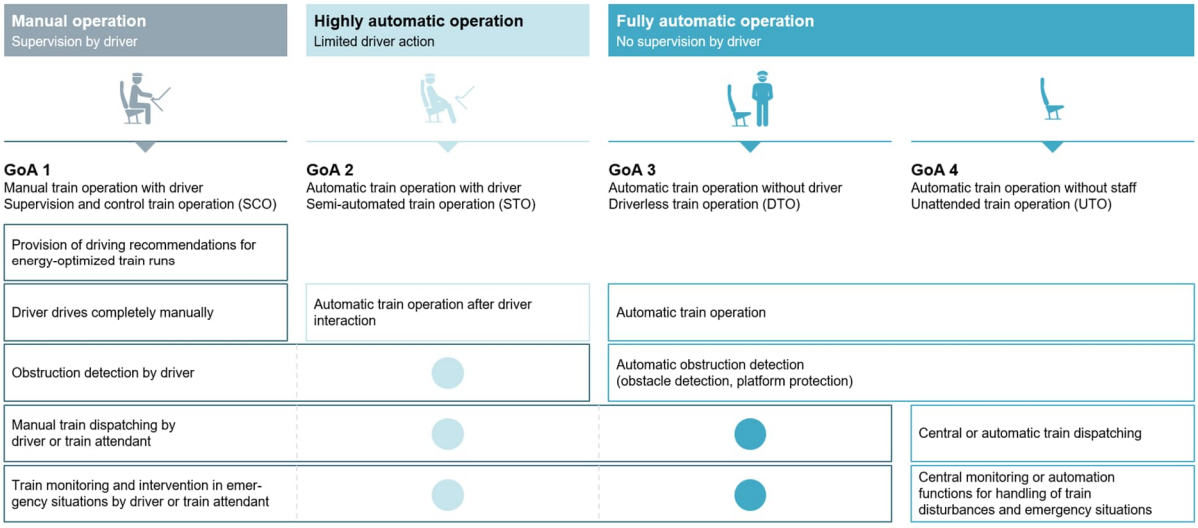


Abbildung 1: Stufen der Automatisierung (GoA) gemäß IEC 62290-1

Fahrerlose Metros und U-Bahnen sind bereits seit über 10 Jahren erfolgreich weltweit im Betrieb. Allerdings operieren diese ausschließlich in kontrollierten, abgeschlossenen Umgebungen. Das vorliegende Projekt zielt auf das Marktsegment der Regionalzüge ab. Diese operieren in einer offeneren Umgebung, in der Hindernisse jeglicher Art (wie z.B. Personen im Fahrweg oder auf der Schiene liegende Bäume, Erdrutsche, etc.) sicher erkannt werden müssen.

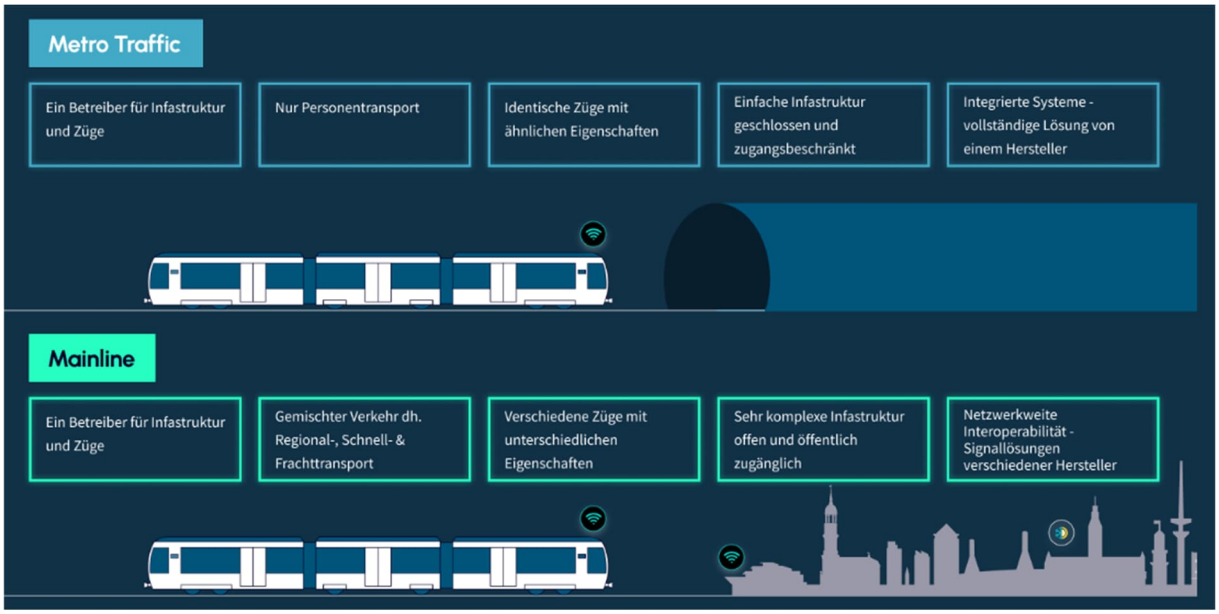


Abbildung 2: Unterschiede zwischen voll-automatisiertem U-Bahnen und Vollbahnen

Für den vollautomatischen Betrieb in offenen Umgebungen (Vollbahn/Mainline) sind aktuell keine Lösungen oder Produkte verfügbar, so dass ein großer Forschungsbedarf auf diesem Technologiefeld besteht.

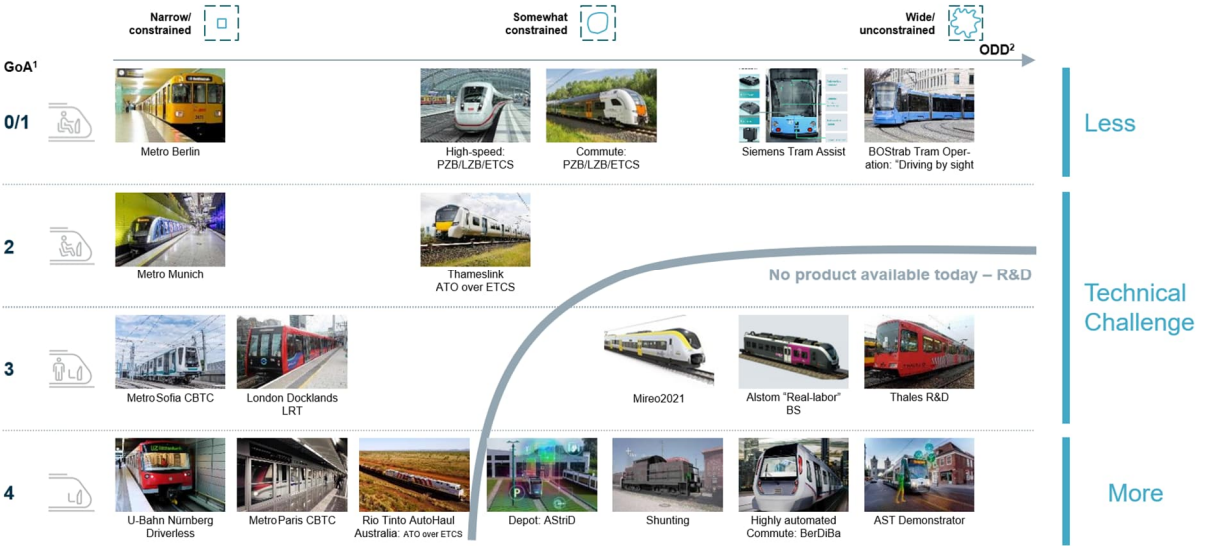


Abbildung 3: Automatisierungsbeispiele im Bezug auf die Betriebsumgebung

Zur Realisierung des vollautomatischen Zugbetriebs sind eine Reihe von Bausteinen/Modulen erforderlich, welche sich zum Teil auf dem Fahrzeug zum anderen Teil an der Landseite befinden. Das vollautomatische Fahren in offenen Umgebungen stützt sich auf die Komponenten, welche für den GoA2-Betrieb (teilautomatisch) erforderlich sind. Darüber hinaus benötigt es weitere neue Systeme und Technologien für das voll-automatisierte Fahren. Die Herausforderung ist die Koordination und Integration der verschiedenen Technologien und Systeme.

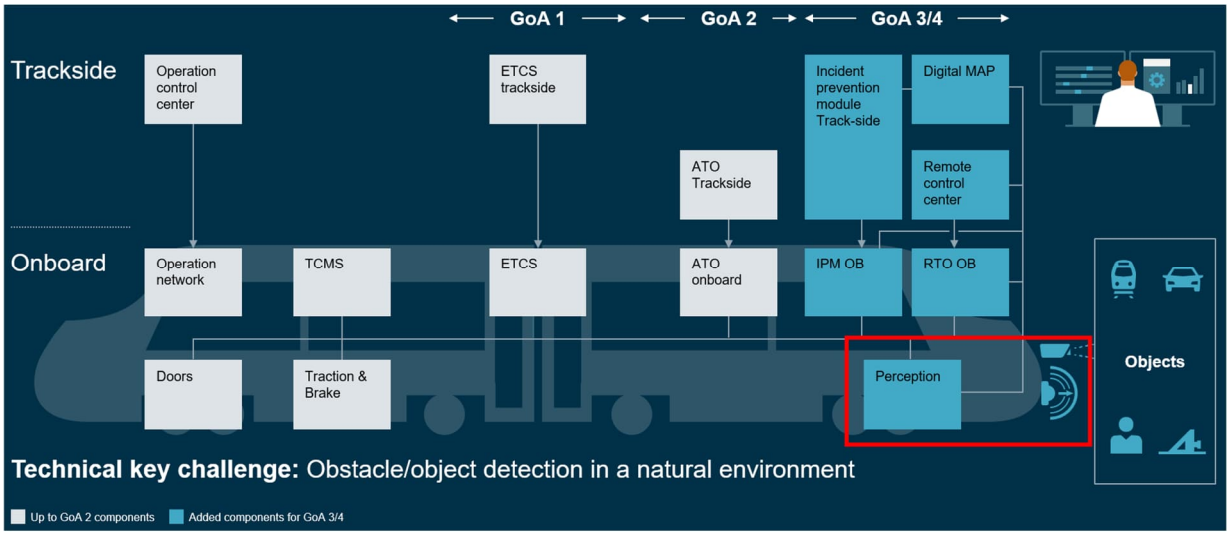


Abbildung 4: Bausteine für den vollautomatischen (GoA3/4-) Betrieb

Eine wesentliche Erweiterung des GoA2-Fahrbetriebs (hoch-automatisierter Betrieb) hin zu einem GoA3/4-Fahrbetrieb (voll-automatisierter Betrieb) stellt die Funktion der voll-automatischen Fahrwegsüberwachung und die damit verbundene Erkennung von Hindernissen im eigenen Fahrweg dar. Diese muss in der Lage sein, den ohne menschliches Eingreifen den Zug sicher zu betreiben, Hindernisse auf der Strecke zu erkennen und die nötige Reaktion einzuleiten.

Die Erkennung von Hindernissen, ebenso wie die Erkennung des Fahrweges, sind geeignete Aufgaben für den Einsatz eines Systems auf Grundlage von KI-basierten Funktionen.

Im Projekt safe.trAIn wurden in einem Konsortium aus Industrie, Forschung, Standardisierungsorganisationen und Gutachtern die technischen Grundlagen für die sichere Anwendung von künstlicher Intelligenz (KI) in einem Perzeptionssystem für fahrerlose Regionalzüge entwickelt. Dazu gehören:

- Analyse bestehender Standards und Regularien
- Entwicklung einer Sicherheitsarchitektur inkl. von (Sicherheits-) Anforderungen und eines Test-Systems (System under Test = SuT)
- Entwicklung von Metriken und Testkriterien zur Verifizierung des Systems
- Entwicklung und Aufbau einer Testumgebung (Virtual Testfield = VTF) und Teststrategie sowie Testkonzept
- Entwicklung eines Sicherheitsnachweis-Konzeptes
- Erstellung eines Konzept-Gutachtens zur entwickelten Vorgehensweise
- Ableitung von Standardisierungspotential von KI-basierten Systemen im Schienenverkehr

Die Ergebnisse des Projekts stellen einen wesentlichen Beitrag zur Weiterentwicklung der Automatisierung im Schienenverkehr dar. Durch die Entwicklung von Methoden zur Sicherheitsbetrachtung von KI-basierten Systemen zur Perzeption können die Voraussetzungen für den Einsatz fahrerloser Regionalzüge geschaffen werden.

Abbildung 5: Übersicht Herausforderungen und Projektziele

1.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Nach dem aktuellen Stand der Technik ist konventionelle Automatisierungstechnik für die Aufgabe der Umfeldwahrnehmung und die Erkennung von Hindernissen nicht ausreichend. Fahrerlose Züge benötigen ein sicheres, zuverlässiges Verständnis der Umgebung, einschließlich einer robusten und sicheren Hinderniserkennung unter hohen Verfügbarkeitsanforderungen. KI birgt hierfür ein großes Potential. Jedoch ist die Entwicklung eines KI-basierten Perzeptionssystems mit erheblichen Herausforderungen verbunden.

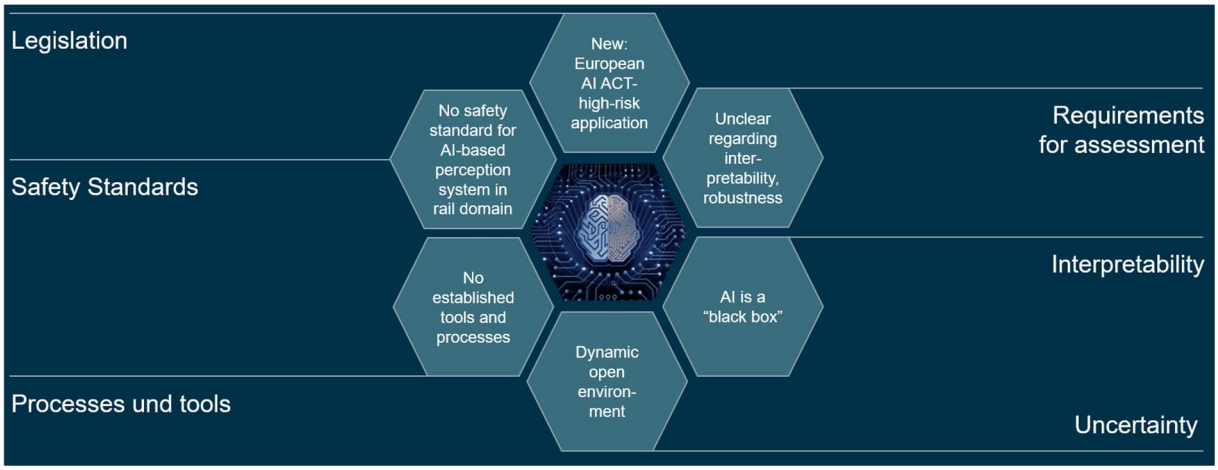


Abbildung 6: Herausforderungen bei der Nutzung von KI in sicherheitskritischen Anwendungsgebieten

- Im sicherheitskritischen Umfeld wie dem Schienenverkehr müssen KI-Systeme hohe Anforderungen an Zuverlässigkeit und Sicherheit erfüllen.
- Die komplexen und dynamischen Umgebungen erfordern eine zuverlässige Erkennung und Klassifizierung von Hindernissen unter allen Bedingungen.
- Es besteht die Herausforderung, die Entscheidungsprozesse von KI-Systemen (von Natur aus „Black-Box"- Charakter) transparent und nachvollziehbar zu machen, was für die Zertifizierung und Akzeptanz durch Regulierungsbehörden, Stakeholder und die Öffentlichkeit unerlässlich ist.
- Zudem müssen KI-Systeme robust gegenüber Störungen und Angriffen sein, um eine ausfallsichere Funktion zu gewährleisten.
- Es existieren keine Normen und etablierten Prozesse für die Entwicklung eines KI-basierten Perzeptionssystem.
- Es existieren keine Vorgaben für die Begutachtung.

Die Bewältigung dieser Herausforderungen erfordert einen strukturierten, systematischen Ansatz, robuste und transparente KI-Systeme und spezialisierte Methoden, die die Lücke zwischen den Fortschritten der KI-Technologie und etablierten sicherheitstechnischen Praktiken schließen.

Der Ausgangspunkt des Projektes stellte der Technologiereifegrad (TRL) 2 bzw. 3 dar (Beschreibung der Anwendung bzw. Nachweis der Funktionalität), welche während des Projektes auf die höheren Stufen gehoben werden konnte. Ziel dabei war, eine hohe Anwendungsnähe, aber keine Produktreife zu erlangen, was einem TRL 4 (Labor) – TRL 6 (Prototyp) entspricht. Die Technologien sollten so weit entwickelt sein, dass ein Konzept eines Sicherheitsnachweises im Projekt erstellt werden kann. Da reale Testfahrten im laufenden Bahnbetrieb nur eingeschränkt und mit hohem Aufwand möglich sind, sollte der Nachweis im virtuellen Testfeld in einer simulierten Einsatzumgebung erfolgen. Ein vollständiger Sicherheitsnachweis, bei dem alle notwendigen Eigenschaften nachgewiesen werden können, wird nicht angestrebt, aber das prinzipielle Vorgehen sollte ersichtlich sein.

Die Entwicklung solcher komplexen KI-Systeme und deren Sicherheitsargumentation erfordert eine enge Interaktion zwischen KI-Experten, Ingenieuren aus dem Schienenfahrzeug-Bereich und Sicherheitsspezialisten. Hierbei war die enge Zusammenarbeit mit den Konsortialpartnern aus Industrie und Forschung ein wesentlicher Faktor. Durch den Austausch von Wissen aus den unterschiedlichen Bereichen (wie z. B. Automotive) konnte das Projekt effizient vorangetrieben werden.

1.3 Planung und Ablauf des Vorhabens

Das Projekt safe.trAIn wurde nach einem detaillierten Meilensteinplan durchgeführt, der die verschiedenen Phasen und Ziele des Vorhabens strukturiert. Die ursprüngliche Planung sah eine Laufzeit von 36 Monaten vor, unterteilt in mehrere Arbeitspakete (AP1 – AP6), die jeweils spezifische Aufgaben umfassten.

1. **Anforderungen an die Sicherheitsverifizierung:** Analyse existierender Normen und Regularien
2. **Methoden und Tools für die Verifizierung:** Entwicklung von Methoden (Metriken) zur Überprüfung des Systems
3. **Sicherheitsarchitektur für KI-basierte Hinderniserkennung:** Entwicklung einer Sicherheitsarchitektur, eines System-under-Test (SuT) und eines Sicherheitsnachweis-Konzepts
4. **Virtuelles Testfeld und Sicherheitsbewertung:** Entwicklung einer virtuellen Test-Umgebung inkl. Testkonzept und -strategie und Begutachtung des Sicherheitsnachweis-Konzepts.
5. **Standardisierung und Verwertung:** Ableitung von Standardisierungspotential für KI-basierten Systemen im Schienenverkehr
6. **Projektmanagement:** Überwachung des Projektfortschritts und Koordination der Arbeiten.

Aufgrund der Komplexität des Projekts und der Verzögerung der Bereitstellung der nötigen Test-Daten zur Validierung der entwickelten Ansätze, wurde das Vorhaben um 3 Monate auf in Summe 39 Monate verlängert. Ein weiterer Grund für diese Verlängerung waren die Arbeiten an der DIN DKE Spec 99004 „Specification of Operational Design Domain in Rail“, welche mehr Zeit erforderten, um ein Standardisierungsdokument mit inhaltlich hoher Qualität und den notwendigen Feedback-Schleifen mit den involvierten Partnern zu erzielen zu können.

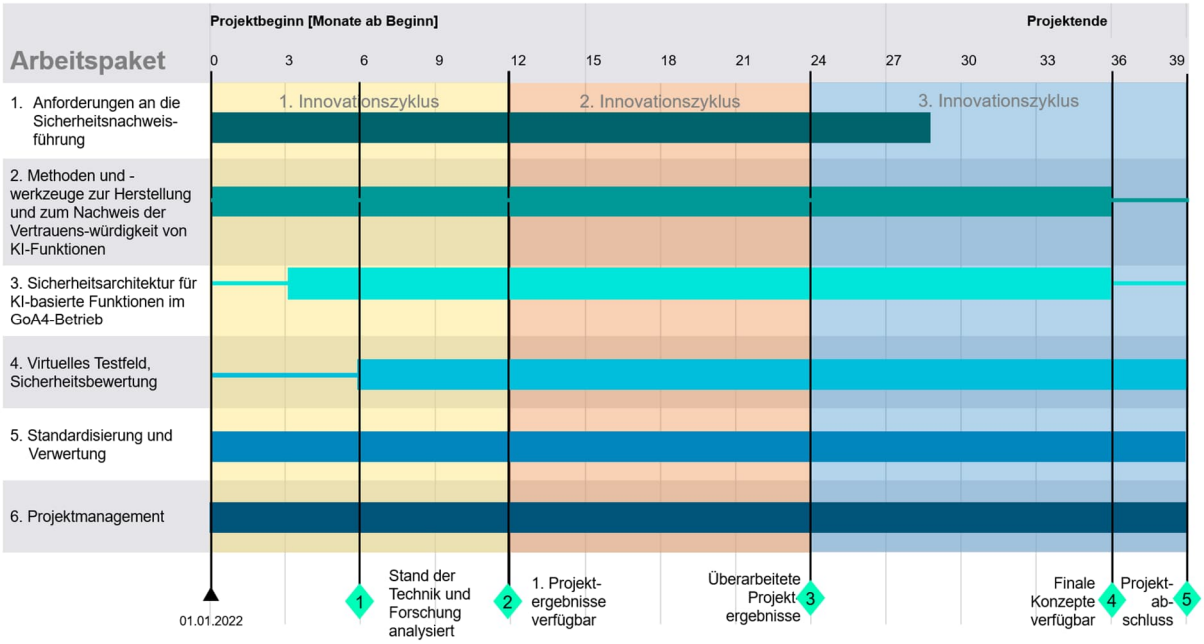


Abbildung 7: Zeitlicher Projektablauf

Der Ablauf des Projekts folgte einem iterativen Ansatz, bei dem die Ergebnisse aller Arbeitsströme und aller Projektpartner als Grundlage für die nächste Iteration dienten. Zur Verfeinerung der 3 ursprünglich geplanten Iterationszyklen wurden sogenannte Minimal Viable Products (= MVPs) eingeführt, welche zunächst den ersten „Durchstich“ ermöglichten und im Anschluss der stetigen Verfeinerung der Ergebnisse dienten.

Abgeleitet von den aktuellen Schutzzielen des Bahn-Sektors konnten zwei Use Cases für das Projekt definiert werden (siehe dazu auch Kapitel 2.1.3.1):

- Die Erkennung von Personen auf dem Gleis.
- Die Erkennung von großen Objekten auf dem Gleis, die das Fahrzeug und somit die Insassen gefährden können.

Aufgrund der Verfügbarkeit von Daten und Wissensbasen (hauptsächlich aus dem Automotive-Sektor) wurde sich zunächst im Rahmen der MVP 1.x-Reihe (MVP 1.1 – MVP 1.3) auf den Use Case „Personen auf dem Gleis“ fokussiert. Im Anschluss daran wurde sich in der zweiten Projekthälfte mit dem zweiten Use Case „Große Hindernisse auf dem Gleis“ im Zuge der MVP2.x-Reihe (MVP 2.0 – MVP 2.2) befasst.

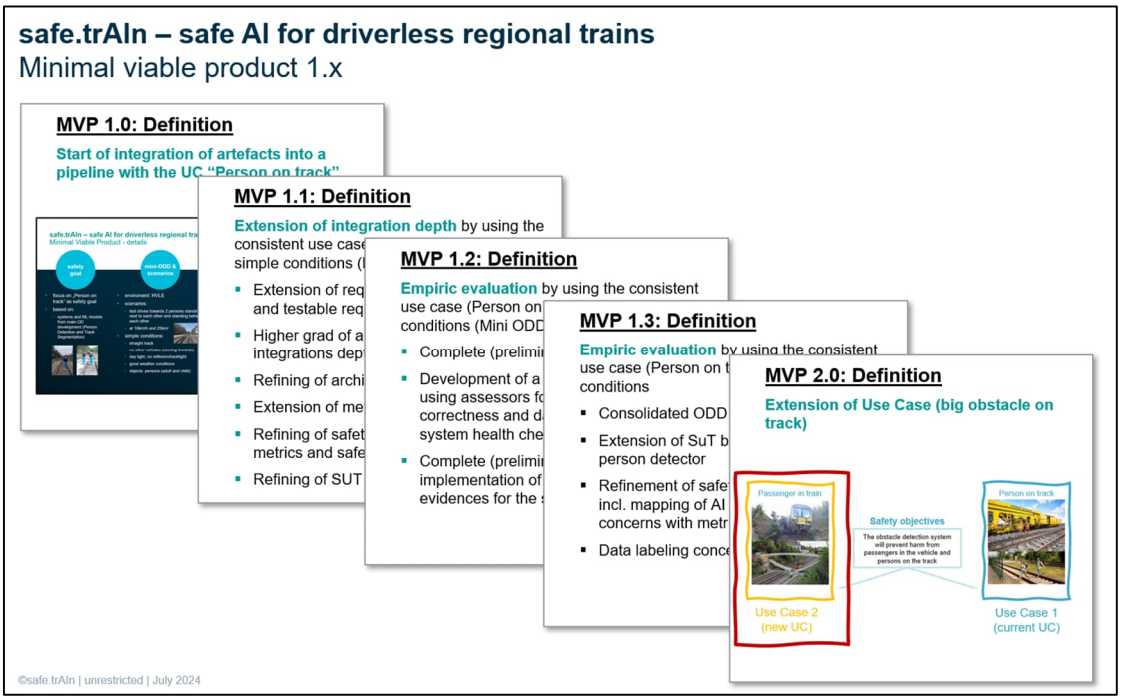


Abbildung 8: Darstellung der MVPs im Projektverlauf (I)

- MVP 1.0: Im ersten MVP wurde sich auf den Use Case „Erkennung von Personen auf dem Gleis“ unter idealen Umgebungsbedingungen geeinigt. Damit konnte der erste Durchstich unter Einbindung aller Arbeitsströme und Projektpartner erlangt werden.

- MVP 1.1: In der nächsten Iterationsphase wurde für den definierten Use Case die Integrationstiefe erweitert. Dabei wurden beispielsweise erste Ansätze zur Sensorfusion in das System under Test integriert, sowie die Anforderungen, Architektur und Metriken weiter ausgearbeitet. Eine wichtige Rolle spielten hier auch die Sammlung KI-relevanter Sicherheitsbedenken („Landscape of AI Safety Concerns“), welche die Basis für die Sicherheitsnachweisführung darstellen.

- MVP 1.2: In diesem MVP konnte eine vorläufige Implementierung des SuT realisiert werden. Hier wurde das erste Konzept von Monitoren in der Architektur entwickelt, welches ein wesentlicher Bestandteil der Sicherheits-Architektur darstellt.

- MVP 1.3: In dieser Iteration lag der Fokus auf der empirischen Validierung des vorläufigen System under Test, welches um weitere Detektoren und Fusionsansätze ergänzt wurde. Darüber hinaus wurde eine konsolidierte Beschreibung der Operational Design Domain (ODD) in Phython erstellt sowie die KI-relevanten Sicherheitsbedenken

(„Landscape of AI Safety Concerns“) verfeinert. Auch ein Konzept zum einheitlichen Labeln der Daten wurde erarbeitet.

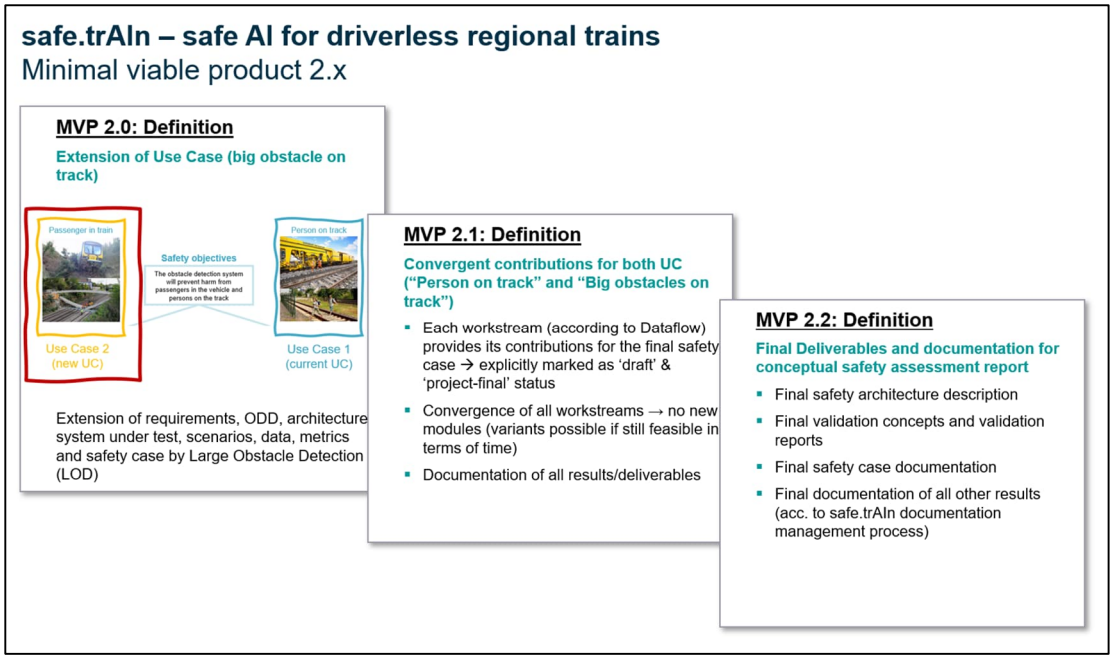


Abbildung 9: Darstellung der MVPs im Projektverlauf (II)

Parallel zu den Arbeiten an MVP 1.3 wurde mit den Vorbereitungen für den zweiten Use Case „Die Erkennung von großen Objekten auf dem Gleis“ gestartet. Dabei wurde sich mit der Spezifizierung der „großen Objekte“ (Wie kann man große Objekte definieren?), der Verfügbarkeit von möglichen Detektoren (Welche Detektoren können genutzt werden?) und die Beschaffung von Daten (Welche Daten können genutzt werden?) befasst.

- MVP 2.0: In diesem Iterationszyklus wurde das System basierend auf den Vorarbeiten um den Use Case „große Objekte“ erweitert. Dies betraf die Anforderungen, die Architektur, das System under Test, die Testszenarien, die Metriken und die Sicherheitsnachweisführung

- MVP 2.1: In der vorletzten Iteration wurde sich darauf fokussiert, die erlangten Ergebnisse aus beiden Use Cases zu konvergieren. Weiterhin wurde ein projekt-internes Dokumentations-Management mit entsprechendem Freigabe-Prozess und Nachverfolgbarkeit aufgesetzt. Dies ist insbesondere für die Ausarbeitung des Sicherheitsnachweises und die anschließende Begutachtung essenziell.

- MVP 2.2: Im letzten MVP wurden alle Ergebnisse final zusammengeführt und die Dokumentation aller Arbeitsströme projekt-final abgeschlossen, so dass alle Artefakte für eine abschließende Begutachtung des Sicherheitsnachweis-Konzeptes vorlagen.

1.4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Das Projekt safe.trAIn basiert auf dem aktuellen Stand der Wissenschaft und Technik im Bereich der künstlichen Intelligenz, Bilderkennung (Computer Vision) und Sensorik für die Umfeldwahrnehmung / Perzeption. Die Entwicklung eines KI-basierten Perzeptionssystems für einen fahrerlosen Regionalzug orientiert sich an bewährten Technologien und Methoden, die in anderen Industrien wie der Automobilindustrie bereits erfolgreich erprobt wurden. Diese wurden für die spezifischen Anforderungen des Schienenverkehrs adaptiert. Dazu gehören:

- Neuronale Netze (maschinelles Lernen / Deep Learning):
  Diese bilden die Grundlage für die KI-basierte Hinderniserkennung.

- Sensorfusion:
  Die Kombination verschiedener Sensoren wie Lidar und Kameras ermöglicht eine zuverlässige Umfelderkennung, wie sie auch in der Automobilindustrie zum Einsatz kommen.

- Sicherheitsarchitektur:
  Die Entwicklung einer robusten Sicherheitsarchitektur basiert auf Standards und Richtlinien der Eisenbahnindustrie (V-Modell) und dem Ansatz des Model based System Engineerings. Darüber hinaus stützt sich die Highlevel Architektur von safe.trAIn auf das seit Jahren etablierte JDL-Pattern (Joint Directors of Laboratories) [Steinberg e.a., Revising the JDL model, Proceedings of the SPIE, 1998].

- Sicherheitsnachweis
  Bei der Konzipierung eines Sicherheitsnachweis für ein KI-basiertes Perzeptionssystem wurde sich an den für die Eisenbahnindustrie geltenden Sicherheits-Normen (insb. EN 50129) orientiert. Diese aber an die zusätzlichen Bedürfnisse durch KI angepasst.

Darüber hinaus stützt sich das Projekt auf viele verschiedene Erfahrungen und Vorarbeiten, die Siemens Mobility (SMO) in der Vergangenheit gemacht hat. Bereits seit mehreren Jahren beschäftigt sich Siemens Mobility mit dem fahrerlosen Fahren in offenen Umgebungen, ist in verschiedensten Forschungsprojekten dazu tätig und konnte bereits mehrfach Ergebnisse präsentieren, wie bei der Autonomen Tram Potsdam 2018, der Autonomen Straßenbahn im

Depot (AStriD) und im BerDiBa-Projekt (Berliner Digitaler Bahnbetrieb). Die vorgenannten Forschungsaktivitäten ebenso wie die beiden durch das Deutsche Zentrum für Schienenverkehrsforschung (DZSF) und Eisenbahnbundesamt (EBA) beauftragten Projekten ATO Sense und ATO Risk lieferten wichtige Technologien und Erkenntnisse, auf welche in safe.trAIn aufgesetzt werden konnte.

- Autonome Tram Potsdam: In dem Projekt hat Siemens Mobility die erste weltweit voll-automatisch fahrende Straßenbahn in Potsdam entwickelt, welche sowohl andere Verkehrsteilnehmer (Fußgänger, Fahrzeuge) mittels eingebauter Sensorik erfassen konnte als auch zuverlässig ihre eigene Position im Schienennetz bestimmen konnte. Darüber hinaus wurden Straßenbahnsignale erkannt. Dies wurde auf der Innotrans 2018 vor einem weltweiten Publikum demonstriert.
-

Abbildung 10: Autonome Tram Potsdam

- AStriD (Autonome Straßenbahn im Depot): Im Rahmen des vom BMVI geförderten Projekts AStriD wurde von Siemens Mobility ein Konzept für einen digitalen Betriebshof mit dem Schwerpunkt der datentechnischen Vernetzung aller an der Automatisierung beteiligten Komponenten (Steuerungssoftware, Anlagen und Fahrzeuge) erarbeitet. Dazu wurde die Straßenbahn mit einer sensorbasierten Umfelderfassung, einer digitalen Karte, den zu erforschenden Lokalisierungstechnologien, sowie den notwendigen Steuerungssystemen ausgerüstet.

Abbildung 11: Autonome Straßenbahn im Depot (AStriD)

- BerDiBa: Dies ist ein vom Land Berlin (ProFIT-Programm, teils EU-kofinanziert) gefördertes Verbundprojekt, koordiniert von Siemens Mobility. Es erforscht und testet unter realen Bedingungen neue Technologien für den automatisierten Bahnbetrieb im Nahverkehr. Das Konsortium aus 12 Partnern (Siemens, Televic GSP, Teraki, Neurocat, Fraunhofer FOKUS/HHI, DFKI, TU Berlin, Zuse-Institut, etc.) adressiert drei Hauptaspekte: fahrerlose Züge, automatisierte Instandhaltung und Teleoperation (Fernsteuerung) für Fälle, in denen die Automatisierung an ihre Grenzen stößt (z.B. Unwetter). Zentral ist eine hochperformante Umfeldsensorik am und im Zug (Kamera, Lidar, Radar) zur Erkennung von Hindernissen auf der Strecke sowie zur Überwachung von Fahrgästen und Infrastruktur. Ein Schlüsselkonzept ist der Digitale Zwilling des Bahnbetriebs: BerDiBa integriert kontinuierlich Sensordaten, um ein virtuelles Abbild des gesamten Systems (Züge, Innenräume, Umwelt, Infrastruktur) zu erzeugen. KI- und Computer-Vision-Methoden spielen dabei eine große Rolle, z.B. für robuste Objekterkennung in Echtzeit und erklärbare KI (XAI), sodass die Ergebnisse nachvollziehbar bleiben. Die entwickelten KI-Modelle werden sofort im realen Betriebsumfeld erprobt (z.B. auf Berliner S-Bahn-Strecken), um ihre Tauglichkeit nachzuweisen. BerDiBa ergänzt safe.trAIn insofern, als hier die praktische Umsetzung im städtischen Raum getestet wird, während safe.trAIn stärker methodisch-konzeptionell vorgeht.

In den beiden durch das Deutsche Zentrum für Schienenverkehrsforschung (DZSF) und Eisenbahnbundesamt (EBA) beauftragten Projekten wurden Ansätze zur Erfassung der physikalischen Leistungsfähigkeit eines Triebfahrzeugführers (Tfs) als Indikation für den

Status Quo (ATO Sense) und zur Risikobewertung mittels Risikoakzeptanzkriterien für den Zugbetrieb ohne menschlichen Tf (ATO Risk) entwickelt.

- ATO-Sense („Sensors and Logik“): Dieses Projekt untersuchte experimentell die Leistungsfähigkeit menschlicher Lokführer in der visuellen Wahrnehmung, um daraus die erforderlichen Spezifikationen für Sensorsysteme eines automatisierten Zuges abzuleiten. In Simulatorstudien (TU Berlin, DLR) wurde gemessen, wie zuverlässig und schnell Menschen Hindernisse erkennen, wie Fehler verteilt sind, und welche Rolle Faktoren wie Aufmerksamkeit und Ermüdung spielen. Das Ergebnis war ein Modell der menschlichen Wahrnehmungsleistung, das als Benchmark dient: Ein ATO-System muss mindestens so gut wie ein menschlicher Fahrer sein, um akzeptiert zu werden. ATO-Sense wurde von der TU Berlin (Fachgebiet Bahnbetrieb) koordiniert mit Partnern DB Systemtechnik, DLR und Siemens Mobility.

- ATO-Risk („Risiko Akzeptanzkriterien“): Dieses Projekt definierte akzeptable Risikoakzeptanzkriterien für vollautomatisierte Züge. Konkret wurde untersucht, welche Fehlerwahrscheinlichkeit man einem KI-basierten Zug erlauben kann und wie groß das Risiko für Personen/Sachen-Schäden dabei sein dürfte, sodass das Sicherheitsniveau mindestens gleichwertig zu manuell durch einen Tf gesteuerten Zügen bleibt. Es wurden Mindestanforderungen an Sicherheit und IT-Security abgeleitet und ein Abwägungsrahmen zwischen Sicherheit und Innovation entwickelt. Siemens Mobility leitete dieses Projekt mit TU Berlin und TÜV Rheinland als Partnern. Die gewonnenen Kriterien helfen Behörden und Industrie, zukünftige Zulassungsverfahren für vollautomatisierte Züge zu gestalten.

Die in ATO Sense und ATO Risk erlangten Ergebnisse wurden in safe.trAIn aufgegriffen. Insbesondere die Ergebnisse von ATO Risk zur Sicherheitseinstufung haben eine wichtige Rolle gespielt, da sie das in safe.trAIn ermittelte Sicherheitsziel (SIL2) bestätigt haben.

Auf europäischer Ebene war Siemens Mobility in dem Bahn- und Forschungsprogramm Shift2Rail (S2R, 2014–2020) aktiv und ist auch weiterhin im Nachfolgerprojekt ERJU (Europe's Rail Joint Undertaking, ab 2021, R2DATO) involviert.
Im Rahmen von Shift2Rail wurden mehrere relevante Projekte zu ATO und KI durchgeführt. Beispielsweise:

- X2Rail-Reihe: Teilprojekte innerhalb S2R, die u.a. ATO über ETCS, Kommunikation und Moving Block thematisierten. Ergebnisse aus X2Rail-4 flossen z.B. in Konzepte für Fernsteuerung und GoA4-Architekturen ein.

- TAURO (Technologies for Autonomous Rail Operation, 2020–2022): Ein S2R-Projekt unter spanischer Leitung (CAF) mit Industriepartnern (Alstom, Bombardier, DB, DLR, Knorr-Bremse etc.), das Schlüsseltechnologien für vollautmatisierte Züge identifizierte. TAURO untersuchte u.a. KI-basierte Umfeldwahrnehmung („artificial sense“) und schlug ein Konzept vor, wie KI-Sensorsysteme zertifizierbar gemacht werden können. Es wurde eine gemeinsame Trainingsdaten-Plattform (Data Factory) konzipiert und die Nutzung von Lidar/Radar-Landmarken für die Zugortung geprüft. Ferner entwickelten die Partner Ansätze für Fernsteuerung (Remote Operation) im Bahnbetrieb für den Fall von Störungen (als Rückfallebene für den vollautomatisierten Zugbetrieb). Ergebnisse von TAURO sollen in Standards (z.B. IEC 61375 für TCMS) eingebracht werden und den Übergang zu GoA4/ATO erleichtern.

Das Forschungsprojekt „MBPLE4Mobility“, ein Konsortium von Siemens Mobility gemeinsam mit Partnern aus Industrie und Forschung, verfolgte das Ziel, modellbasiertes Systems Engineering (MBSE) mit Product Line Engineering (PLE) zu einem durchgängigen Entwicklungsansatz zu vereinen. Dieser Ansatz wurde für komplexe Steuerungssysteme im Schienenverkehr entwickelt und im Projekt im Kontext des teilautomatisierten Fahrens (GoA2, ATO over ETCS) durchgeführt. Die in dem Projekt entwickelten Methoden insbesondere hinsichtlich Software- und Systemarchitektur sind wichtige Grundlagen für die Architektur-Entwicklung im Projekt safe.trAIn gewesen.

Der im Zuge des Siemens Mobility Deep Learning Factory (DLF) Projektes entwickelte ai.store als Datenmanagementlösung für das Speichern von großen Mengen an Trainings- und Testdaten und KI-Modellen stellte einen essenziellen Baustein im Projekt safe.trAIn dar. Durch das Freischalten der Projektpartner im ai.store, erhielten diese Zugriff auf:

- Rohdaten und vorverarbeitete Daten von Kamera, Lidar und weiterer Sensoren
- Kontextinformationen zu den Daten, z.B. zur Datenerhebung, Annotationen, etc.
- Struktur und Parameter von Machine Learning Modellen wie z.B. Neuronale Netze
- Software, z.B. zum Trainieren von Machine Learning Modellen, Vorverarbeitungsschritten, Datenfusion, etc.

1.5 Zusammenarbeit mit anderen Stellen

Eine Zusammenarbeit mit anderen Stellen über die Projektpartner hinaus fand im Rahmen des Projektes nicht statt. Die Konsortialpartner sind im Folgenden aufgelistet:

- Siemens AG
- Fraunhofer Institut für Intelligente Analyse- und Informationssysteme (IAIS)
- Fraunhofer Institut für Kognitive Systeme (IKS)
- Bridgefield GmbH
- SETLabs GmbH
- BIT Technology Solutions GmbH
- Merantix Momentum GmbH
- ITQ GmbH
- Edge Case Research GmbH
- TÜV Nord AG
- TÜV Süd AG
- TÜV Rheinland AG
- Hochschule Düsseldorf
- Otto-von-Guericke-Universität Magdeburg
- DIN e. V. (Deutsches Institut für Normung)
- VDE e. V. (Verband der Elektrotechnik Elektronik Informationstechnik)

Im Zuge der Standardisierungstätigkeiten im Projekt wurden jedoch vom Deutschen Institut für Normung e. V. (DIN) Anwenderkreis-Meetings mit projekt-externen Fachexperten angesetzt, welche es ermöglichten, die Ergebnisse und Erfahrungen des Projekts mit anderen Akteuren aus dem Bahnsektor, aber auch mit anderen Sektoren und Wissenschaft auszutauschen.
Davon und von den Erkenntnissen im Projektverlauf abgeleitet konnten Standardisierungspotentiale identifiziert werden, welche im Rahmen des Projektes in zwei DIN DKE Spezifikationen mündeten. Die Erarbeitung der beiden DIN DKE Spezifikationen erfolgte gemeinsam mit Siemens Mobility, safe.trAIn Partnern und weiteren externen Stakeholdern aus dem Bahnsektor.

2 Eingehende Darstellung

2.1 Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

Das Projekt safe.trAIn hat die Zuwendungsmittel gezielt eingesetzt, um die im Antrag definierten Themenfelder zu erarbeiten und die dazugehörigen Projektziele zu erreichen.

Im Projektverlauf wurde ein detaillierter Workflow entwickelt, der den strukturierten Entwicklungs- und Testprozess für ein KI-basiertes Perzeptionssystem beschreibt und die Zusammenhänge und Abhängigkeiten der einzelnen Arbeitspakete darstellt.

Der Workflow beginnt mit einer umfangreichen Standardisierungsanalyse, um den Entwicklungsprozess sorgfältig an bestehende und zukünftige internationale Sicherheitsstandards anzupassen. Ein zentraler Bestandteil ist die Etablierung expliziter und messbarer Methoden und Metriken, die eine transparente Validierung sicherheitskritischer KI-Funktionalitäten gewährleisten. Wesentlich für diesen Workflow ist die systematische Definition klarer Sicherheitsanforderungen und Betriebsbedingungen, die auf die explizit definierte Operational Design Domain (ODD) abgestimmt sind. Das strukturierte architekturelle System Design und die prototypischen Implementierungen sind speziell an den Use Case fahrerloser Regionalzug angepasst, um Lösungen zu bieten, die den praktischen Betriebsanforderungen entsprechen. System- und Integrationstests in einer fortschrittlichen virtuellen Simulationsumgebungen (virtuelles Testfeld) bieten eine umfassende Validierung der Systemleistung und -zuverlässigkeit. Darüber hinaus gewährleisten Laufzeitüberwachungslösungen eine kontinuierliche Bewertung der Systemsicherheit während des Betriebs. Diese einzelnen Arbeitsstränge werden abschließend in ein kohärentes und detailliertes Sicherheitsnachweiskonzept integriert (siehe Abbildung 12).

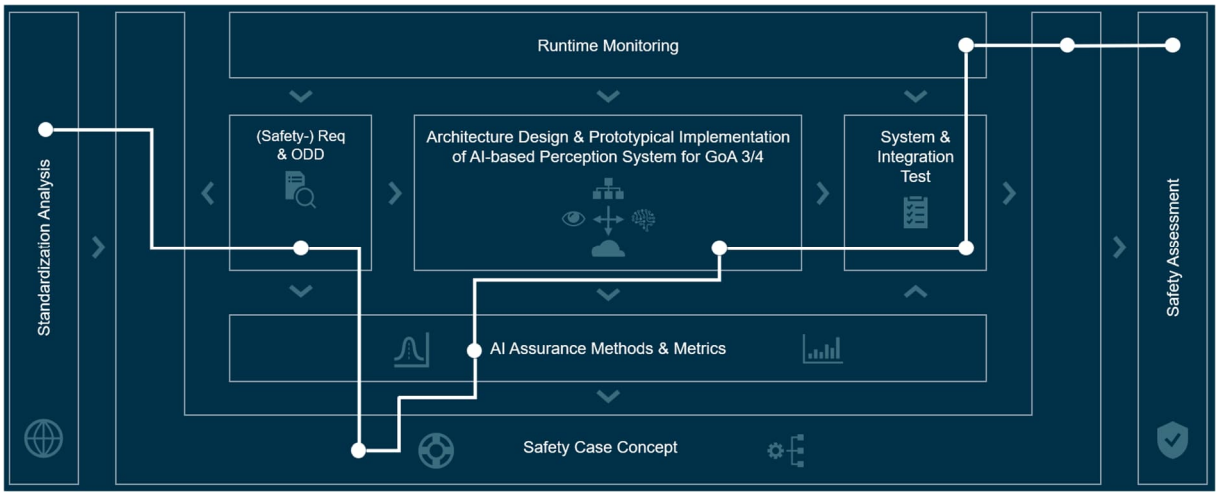


Abbildung 12: safe.trAIn Workflow

Im Nachfolgenden werden die erzielten Ergebnisse je Arbeitspaket (= AP) näher erläutert.

2.1.1 AP 1: Anforderungen an die Sicherheitsnachweisführung

In diesem AP wurden die Anforderungen/Eigenschaften und Akzeptanzkriterien für die Sicherheitsnachweisführung eines KI-basierten Perzeptionssystems eines fahrerlosen Regionalzuges auf Basis bereits bestehender Normen und Standards abgeleitet. Dazu wurde eine ausführliche Recherche und Bestandsaufnahme von bestehenden Normen und Regularien hinsichtlich „KI und funktionaler Sicherheit“ gemeinsam mit vorranging den Projektpartnern von DIN, DKE, Siemens AG und TÜVs durchgeführt.

2.1.1.1 Analyse bestehender Normen und Regelwerke

Es wurden zahlreiche (> 80) Standards, Entwürfe und andere Quellen identifiziert, die Hinweise auf Prozesse, Methoden, Techniken und Werkzeuge enthalten, die auch für sicherheitsrelevante Funktionen bei der Anwendung von KI verwendet werden können. Bei näherer Betrachtung der Anwendbarkeit für den Projekt Use Case „Perzeption eines fahrerlosen Regionalzuges“ konnte festgestellt werden, dass für die meisten von ihnen jedoch nicht genügend zuverlässige Referenzen und empirisches Wissen vorliegen, um klare Empfehlungen zu geben, wie sie für ein bestimmtes SIL angewendet werden können. EN 50716:2023 empfiehlt sogar ausdrücklich, KI nicht zu verwenden.

Es konnten jedoch Standards in anderen Domänen (z. B. Automotive: SOTIF = Safety Of The Intended Functionality) identifiziert werden, welche teils noch in der Entwicklung sind. Ebenso existieren verschiedene Forschungs-Ansätze zu diesem Thema.

Im vorliegenden Projekt ist man zu der Erkenntnis gekommen, dass bestehende Standards für sicherheitsrelevante Zulassungen im Eisenbahnwesen (EN 50128/50129) als Grundlage für KI-basierte Systeme verwendet werden können. Denn Aspekte wie fehlervermeidende Entwicklungsmethoden, fehlererkennende V&V-Maßnahmen und Fehlererkennung & -kontrolle während des Betriebs sind allgemein, auch beim Einsatz von KI-basierten Systemen, erforderlich, aber im Kontext von KI herausfordernder. Im Projekt hat man den Ansatz gewählt, sich auf die im Bahnsektor etablierte Eisenbahn-Norm EN 50129 zu stützen und diese um Methoden und Ansätze des Stand der Technik (State of the Art = SotA) zu ergänzen.

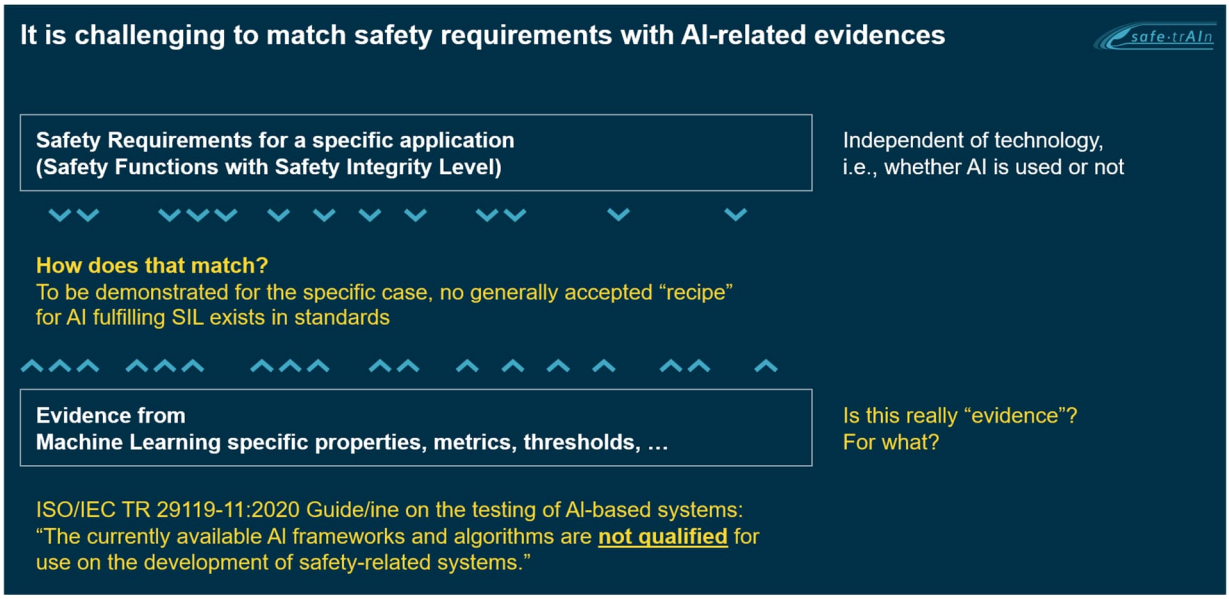


Abbildung 13: Darstellung der Herausforderung: Sicherheitsanforderungen (SIL) mit Evidenzen von KI-Funktionen

2.1.1.2 European AI Act

Im Zuge des Projektes wurde sich auch intensiv mit dem European AI Act und dessen Ausprägungen befasst. Der AI Act schafft einen verbindlichen Rechtsrahmen für den sicheren und vertrauenswürdigen Einsatz von Künstlicher Intelligenz, insbesondere in sicherheitskritischen Bereichen, wozu auch der Schienenverkehr gehört. Für das Projekt safe.trAIn ist der AI Act hochrelevant, da er Anforderungen an Transparenz, Nachvollziehbarkeit und Risikomanagement definiert, die direkt in die Entwicklung und Zulassung der Systeme einfließen müssen. Aufgrund der sektor-übergreifenden, horizontalen

Wirksamkeit des AI Acts sollten die im Projekt entwickelten Methoden und Konzepte AI Act Anforderungen berücksichtigen, um eine Grundlage für eine zukünftige Zulassung eines KI-basierten Perzeptionssystems für einen fahrerlosen Regionalzug zu schaffen. Daher wurde die Entwicklung des European AI Act während des Projektes sehr intensiv beobachtet.

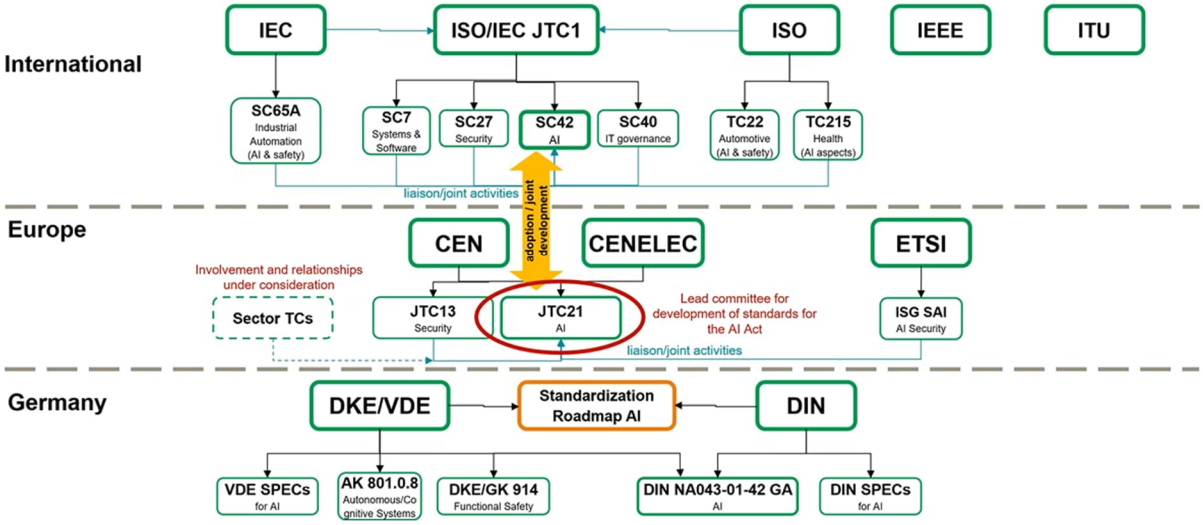


Abbildung 14: Normen-Landschaft für den AI Act

Ein wichtiger Bestandteil in dem Zuge bestand in der tiefen Analyse der ISO/IEC DTR 5469 („Functional safety and AI systems“).

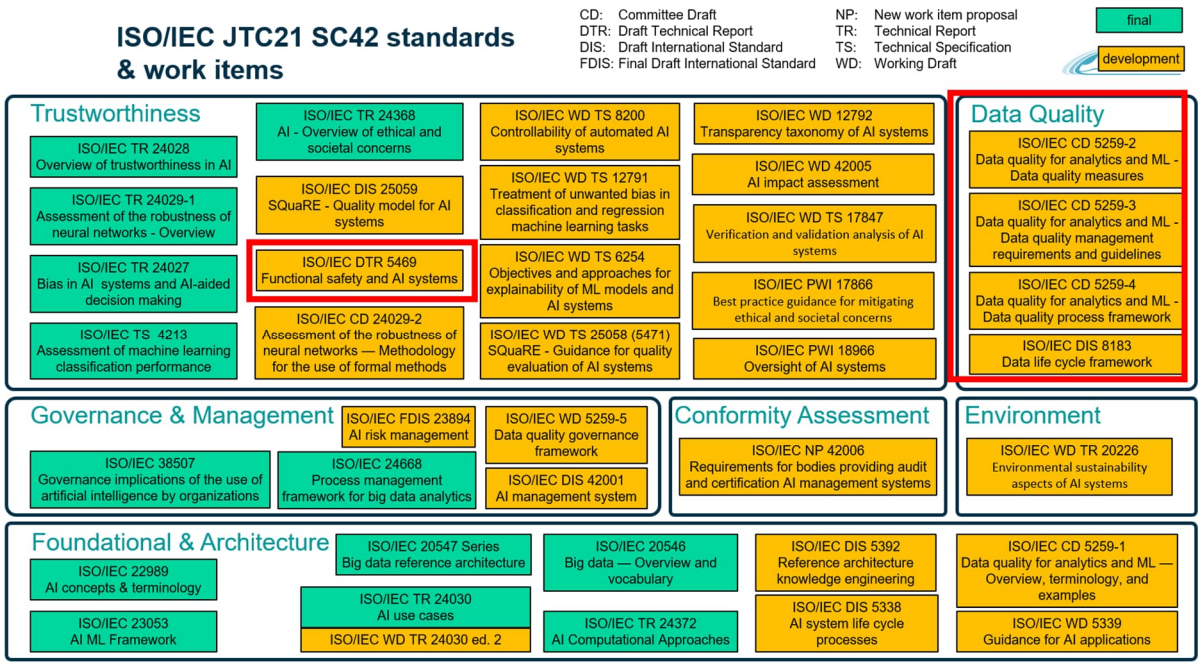


Abbildung 15: JTC21 Normen-Landschaft

Während der AI Act den regulatorischen Rahmen festgelegt hat, liefert die ISO/IEC DTR 5469 die technischen Grundlagen für die Umsetzung der Anforderungen im Kontext der funktionalen Sicherheit. Sie beschreibt, wie KI in sicherheitsrelevanten Funktionen eingesetzt werden kann, welche Risiken dabei bestehen und welche Methoden zur Absicherung geeignet sind. Nachfolgende Abbildungen zeigen die voraussichtliche Einordnung des vollautomatisierten Fahrens mittels KI-basierten Perzeptionssystems gemäß der in der ISO/IEC DTR 5469 vorgegebenen Klassifizierungskriterien.

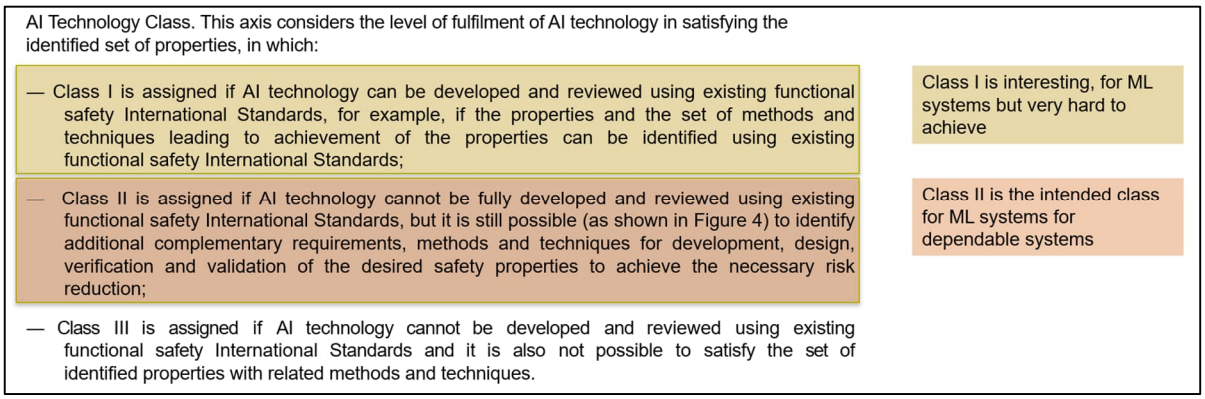


Abbildung 16: KI Technologie Klassen gemäß ISO/IEC DTR 5469

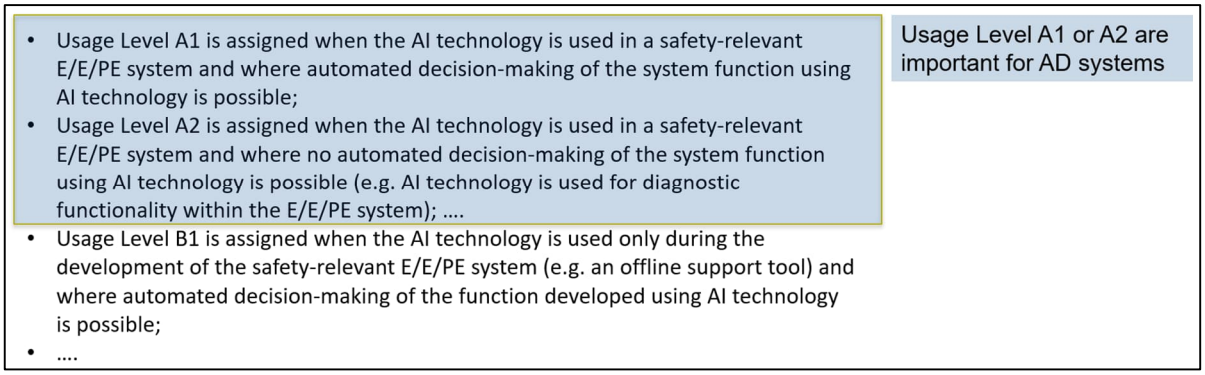


Abbildung 17: Nutzungsebenen gemäß ISO/IEC DTR 5469

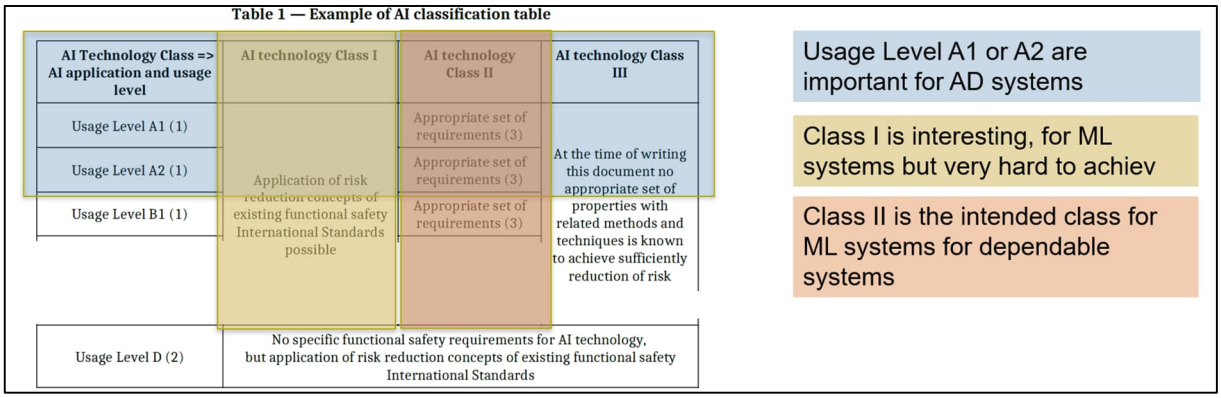


Abbildung 18: exemplarische Klassifizierung gemäß ISO/IEC TR 5469

Mit dem EU AI Act wurde ein Rechtsrahmen geschaffen, der die Nutzung von KI unter speziellen Vorschriften auch für sicherheitsrelevante Anwendungen ermöglichen soll. In Automotive Standards wie der ISO/PAS 8800 („Road vehicles — Safety and artificial intelligence“) wird auf die ISO DTR 5469 verwiesen. Die ISO DTR 5469 sieht eine große Herausforderung für den Einsatz von KI-Technologie bei der Implementierung funktionaler Sicherheitssysteme, wenn die durch maschinelles Lernen abgeleiteten Modellparameter zu komplex sind, um verstanden, analysiert und verifiziert zu werden.

Daraus können folgende Schlüsse für die Entwicklung eines KI-basierten Perzeptionssystems gezogen werden:

- Die Probleme, die sich aus der Nutzung komplexer Machine Learning (ML)-Systeme ergeben können, müssen tiefer verstanden werden. Darüber hinaus muss nach Lösungen gesucht werden, um die Komplexität, die vom ML-/KI-System verwaltet wird, zu reduzieren.
- Die „Black Box“ muss aufgelöst, und untersucht werden. Das komplexe Problem muss in kleinere Teile aufgeteilt werden, die separat entwickelt, verifiziert und validiert werden können, wobei so weit wie möglich die „traditionelle“ Sicherheit angewendet wird,
- Die Komplexität, die in der ML-Komponente gehandhabt wird, muss reduziert werden, um ausreichende Beweise dafür zu liefern, dass das System für den Einsatz in Hochrisikofunktionen ausreichend vertrauenswürdig ist.

Ein weiteres Thema, das im Zuge der Analyse bestehender Standards und des AI Acts, intensiv im Projekt beleuchtet wurde, war das Thema Daten-Qualität. Die ISO/IEC 5259-Serie identifiziert Maßnahmen zur Datenqualität, Anforderungen an das Datenqualitätsmanagement und einen repräsentativen Prozess zur Verwaltung der Datenqualität über den gesamten Datenlebenszyklus.

Die Datenqualität ist entscheidend für die Qualität von ML-Systemen: In Anwendungen des maschinellen Lernens hat die Datenqualität einen großen Einfluss auf die Qualität der Analyseergebnisse und die Leistung des ML-Modells. Die Verwendung von Daten, die nicht den Anforderungen für einen bestimmten Zweck entsprechen, kann zu Modellen von geringer Qualität führen, die ungenau und anfällig für Fehler sind. Das Verstehen, Messen und Verwalten der Merkmale der Datenqualität ist daher sehr wichtig.

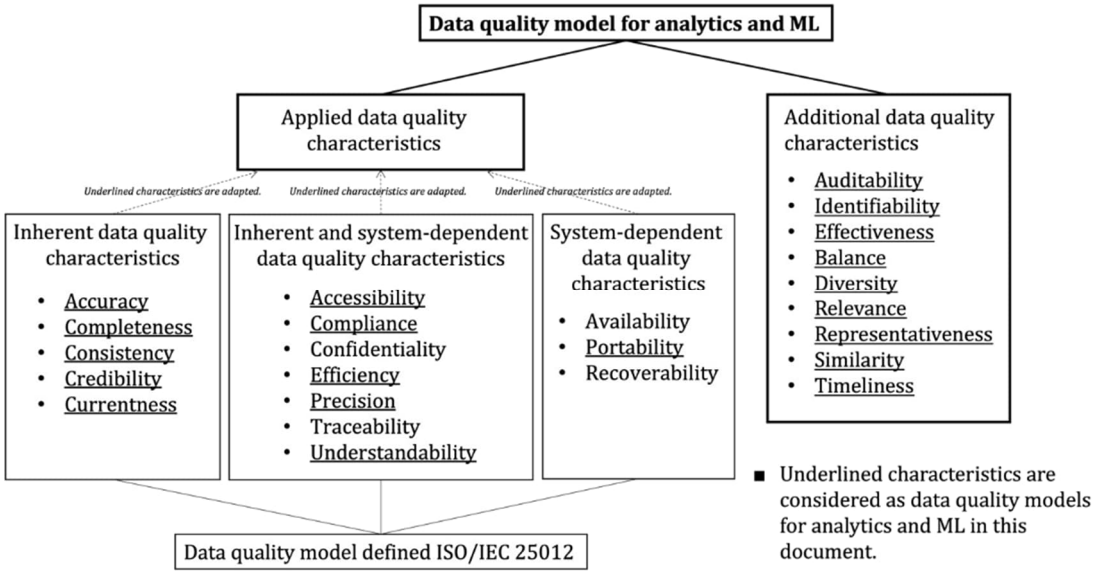


Abbildung 19: Datenqualität für ML-Modelle gemäß IEC/ ISO/IEC 5259-Serie [ISO/IEC WD 5259-2:202X(X) Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures]

Es ist essenziell, zu Beginn und während des Entwicklungsprozesses (in den jeweiligen Entwicklungsphasen) zu bestimmen, welche Datenqualitätsmerkmale in welchem Umfang für die Performance und Sicherheit des KI-Systems entscheidend sind.

	single sensor	sensor-"only" fusion	sensor-map-fusion	single map
currentness	✓	✓	✓	✗
resolution	✓	✓	✓	✓
elimination of random errors	✗	✓	✓	~
redundancy	✗	✓	✓	✗
hypothesis verification	✗	~	✓	✗
field of view	limited to vehicle position and environmental conditions			global
cost quantities	per sensor per vehicle			per content, must be updated regularly
type of information	specialized to sensor			can contain various information

Abbildung 20: Aspekte der Datenqualität während der Daten-Generierung

2.1.1.3 Akzeptanzkriterien für die Sicherheitsfunktion

Weiterhin sollten in diesem Arbeitspakete die Akzeptanzkriterien für die Sicherheitsfunktion abgeleitet werden.

Analysen haben gezeigt, dass Regionalzüge sehr selten gefährdenden Hindernissen begegnen (< 1 Jahr). Daher wurde im Projekt der Ansatz der PFD (= Probability of Failure in Demand, Versagenswahrscheinlichkeit im Bedarfsfall) als Akzeptanzkriterium gewählt. Mittels entsprechender Risikoanalysen mit Ereignisbäumen, welche im Projekt von Sicherheits-Experten durchgeführt wurden, konnte das quantitative Sicherheitsziel von 1% für die PFD abgeleitet werden. Dies entspricht einem Sicherheitsintegritätslevel (SIL) 2, was sich mit den Erkenntnissen aus ATO Risk deckt.

PFD = 1 – Recall

Recall = True Positive / (True Positive + False Negative)

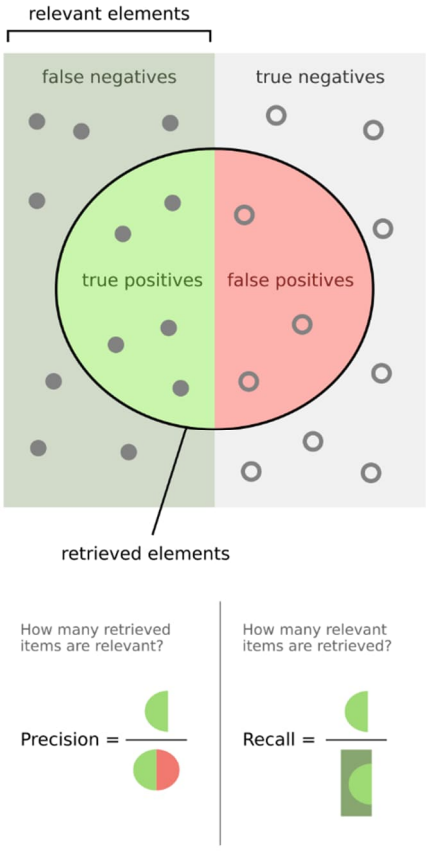


Abbildung 21: Darstellung Recall [Walber - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36926283]

Somit konnte im Projekt die PFD von 1% als grundlegende Sicherheitsanforderung an eine voll-automatische Fahrwegsüberwachung definiert werden. Dies bildet die Grundlage für die Definition weiterer Anforderungen an ein solches System in AP3 (siehe Kapitel 2.1.3.1.)

2.1.2 AP 2: Prüfmethoden und -werkzeuge zum Nachweis der Vertrauenswürdigkeit von KI-Methoden

Auf Basis der im Arbeitspaket 1 identifizierten Eigenschaften und Akzeptanzkriterien hat Arbeitspaket 2 sich mit der Identifikation und Entwicklung von Messgrößen, sogenannten Metriken oder Sicherheitsperformanz-Indikatoren (SPI), zur Validierung und zum Nachweis der Sicherheit und Vertrauenswürdigkeit von KI-Algorithmen befasst.

Dazu wurde zunächst der Stand der Technik (State of the Art) analysiert und untersucht, inwieweit bestehende Methoden zur Performanz-Verbesserung sowie Validierung von KI-Algorithmen eingesetzt oder angepasst werden können. Dies wurde in Form von State-of-the-Art-Reports dokumentiert und zusammengefasst.

Darauf aufbauend wurden die Lücken betrachtet sowie die Vorteile und Limitationen einzelner Metriken dokumentiert. Entgegen der ursprünglichen Vorhabensplanung wurden die identifizierten Lücken nicht separat in Form einer Gap-Analyse niedergeschrieben, da die Dokumentation in den State-of-the-Art-Reports für ausreichend erachtet wurde.

State-of-the-Art-Reports:

- Data Quality & Completeness
- Architectural Methods and Combination with Classical Methods
- Robustness
- XAI State of the art report
- Validation
- Runtime Monitoring

2.1.2.1 Konzept zur systematischen Erstellung eines Nachweises - Landscape of AI Safety Concerns (LAISC)

Um die Lücke zwischen KI-basierten Systemen und konventionellen Softwaresystemen zu schließen, war es nötig, die KI bezogenen Bedenken („AI Safety Concerns“) zusammen zu tragen. So entstand die „Landscape of AI Safety Concerns“ (LAISC), eine innovative Methodik zur Auflistung aller KI relevanten Gefährdungen und möglichen Mitigations-Maßnahmen über die einzelnen KI-Lebenszyklusphasen hinweg [*Schnitzer, R.; Kilian, L.; Roessner, S.; Theodorou, K.; & Zillner, S. (2024), "Landscape of AI Safety Concerns: A Methodology to Support Safety Assurance for AI-based Autonomous Systems," 8th International Conference on System Reliability and Safety (ICSRS), 2024*].

Ein AI Safety Concern ist ein KI spezifisches Belangen, das sich negativ auf die Sicherheit eines Systems auswirken kann. Gemeinsam mit den Projektpartnern, insbesondere Siemens AG, Fraunhofer IKS und Siemens Mobility, konnten in weitreichen State-of-the-Art-Analysen 22 AI Safety Concerns identifiziert werden (siehe Abbildung 22).

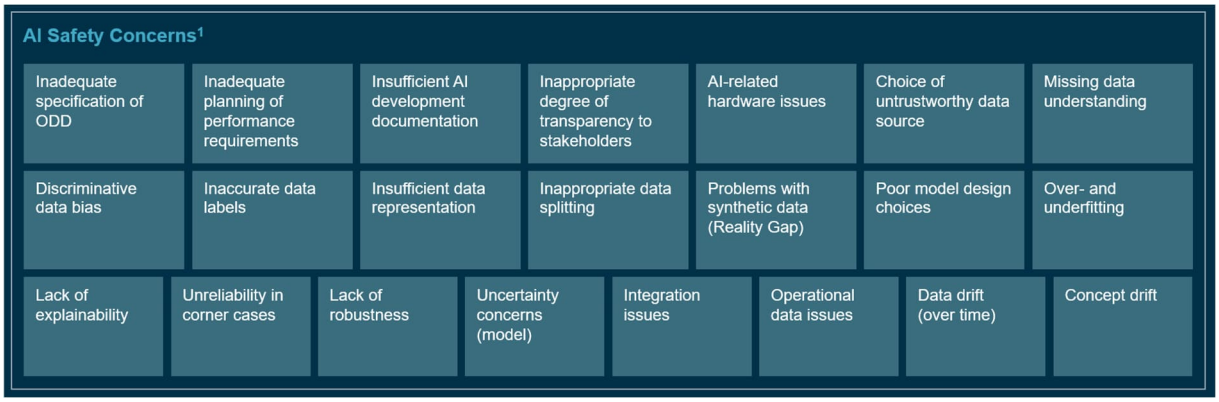


Abbildung 22: Landscape of AI Safety Concerns

Für jedes AI Safety Concern wurden verifizierbare Kriterien abgeleitet („Verifiable Requirements“), mit denen nachgewiesen werden kann, dass eine Gefährdung im System erst gar nicht auftreten oder wie diese wirksam beherrscht werden kann. Mittels Metriken können die nötigen Evidenzen dafür erbracht werden.

Im Projekt wurde ein 4-stufige Vorgehen zur Anwendung der LAISC entwickelt (siehe Abbildung 23).

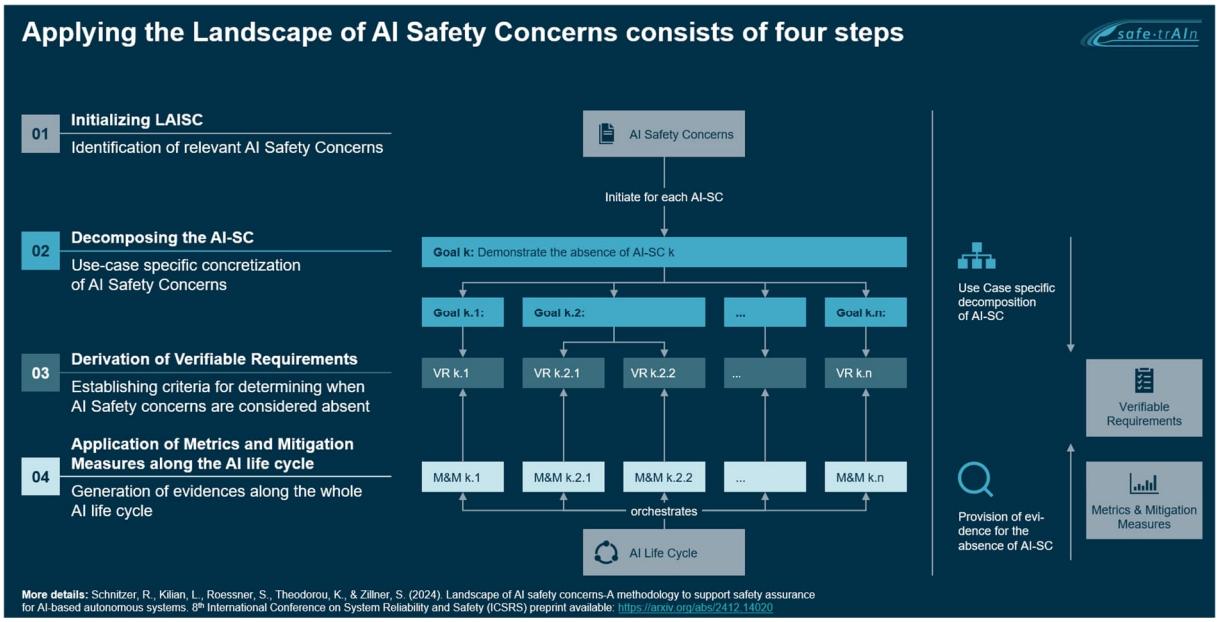


Abbildung 23: 4-stufiger Prozess zur Anwendung der "Landscape of AI safety conerns"

1. Identifizierung der AI Safety Concerns

2. Anwendungsspezifische Zerlegung der identifizierten AI Safety Concerns
   In diesem Schritt werden diese identifizierten AI Safety Concerns spezifisch auf jeden Anwendungsfall zugeschnitten, wobei sie klare und umsetzbare Details geben. Dies ermöglicht präzise und zielgerichtete Sicherheitsmanagementstrategien, die Relevanz und Effektivität sicherstellen.

3. Ableitung überprüfbarer Anforderungen („Verifiable Requirements“) für jedes AI Safety Concerns
   Diese Verifiable Requirements bieten eine Vergleichsmöglichkeit zur Validierung und zum Management jeder AI Safety Concerns

4. Anwendung von Metriken und Mitigations-Maßnahmen über den gesamten KI-Lebenszyklus hinweg
   Dieser Schritt stellt eine umfassende, fortlaufende Validierung der Sicherheits-Maßnahmen sicher und adressiert systematisch jedes identifizierte AI Safety Concern.

Der 4 stufige Prozess in seiner strukturierten Weise gewährleistet Transparenz und Rückverfolgbarkeit.

Am AI Safety Concern „Inaccurate Data Labels” wird der Prozess im Folgenden exemplarisch erklärt.

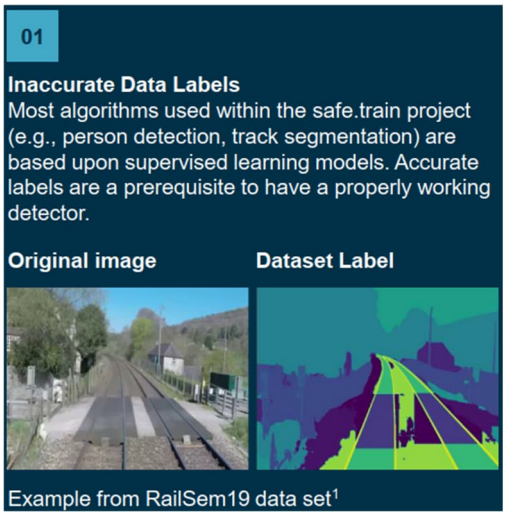


Abbildung 24: AI Safety Concern "Inaccurate Data Labels"

Die Sicherstellung hochwertiger Datenlabels ist sowohl für reale als auch für synthetische Daten von entscheidender Bedeutung, um die Sicherheit eines KI-basierten Perzeptionssystems zu gewährleisten.

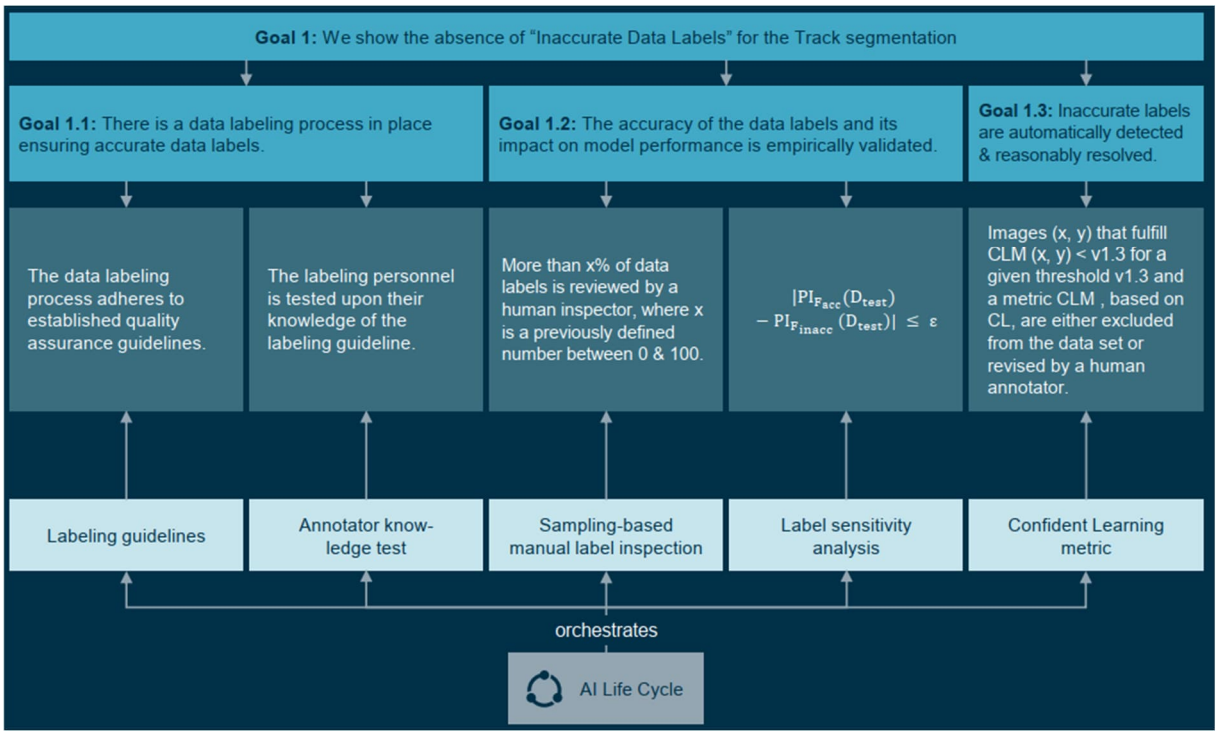


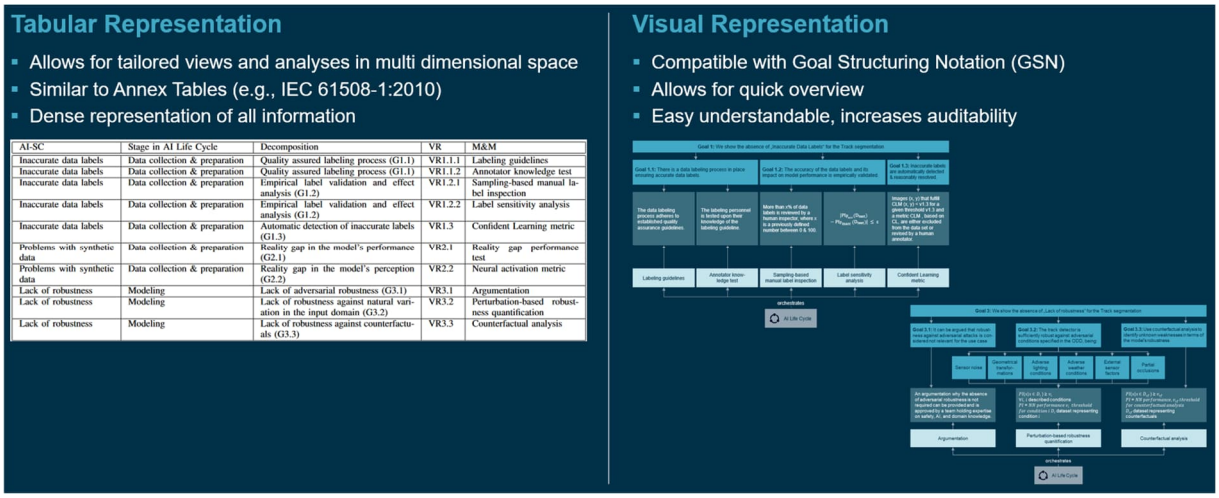
Abbildung 25: Exemplarische Darstellung des 4-stufigen Prozesses der Landscape of AI Safety Concerns anhand "Inaccurate Data Labels"

Abgeleitet von dem AI Safety Concern „Inaccurate Data Labels" wird das Sicherheitsziel („Goal 1") formuliert und in 3 Unterziele (Goal 1.1, Goal 1.2 und Goal 1.3) heruntergebrochen (siehe Abbildung 25). Für jedes dieser Unterziele werden nun die entsprechenden „Verifiable

Requirements“ definiert, für die wiederum Mitigations-Maßnahmen über den KI-Lebenszyklus definiert werden müssen.

Für das Beispiel des „Inaccurate Data Labels“ wurde im Projekt ein Prozess entwickelt, der die Genauigkeit der Datenlabels gewährleistet, die für das Training und die Validierung des KI-Systems verwendet werden können. Es wurde in Zusammenarbeit von Siemens Mobility und Siemens AG ein Labelingkonzept erarbeitet und in einen detaillierten Labeling-Prozess umgesetzt, welcher auch über das Projekt safe.trAIn hinaus bei Siemens Mobility zur Anwendung kommt. Dazu gehört, dass jedes Datenlabeling gründlichen manuellen Inspektionen und automatisierten Validierungsprozessen unterzogen wird, um Ungenauigkeiten zu erkennen und korrigieren zu können. Diese trägt dazu bei, gleichbleibend hochwertige Datenlabelings zu erhalten, die für ein effektives Lernen des KI-Systems (supervised learning) unerlässlich sind. Darüber hinaus wird der Einfluss der Genauigkeit des Datenlabeling auf die Leistung des KI-Modells empirisch untersucht, um sicherzustellen, dass die KI-Modelle die entsprechenden Kriterien erfüllen. Die automatisierte Erkennung und manuelle Behebung von Ungenauigkeiten im Datenlabeling tragen dazu bei, kontinuierlich hohe Standards der Datenqualität aufrechtzuerhalten und somit die Zuverlässigkeit, Robustheit und Sicherheit des Modells erheblich zu verbessern.

Im Projekt wurden 2 Wege aufgezeigt, wie die Landscape of AI Safety Concerns dargestellt werden kann:



AI-SC	Stage in AI Life Cycle	Decomposition	VR	M&M
Inaccurate data labels	Data collection & preparation	Quality assured labeling process (G1.1)	VR1.1.1	Labeling guidelines
Inaccurate data labels	Data collection & preparation	Quality assured labeling process (G1.1)	VR1.1.2	Annotator knowledge test
Inaccurate data labels	Data collection & preparation	Empirical label validation and effect analysis (G1.2)	VR1.2.1	Sampling-based manual label inspection
Inaccurate data labels	Data collection & preparation	Empirical label validation and effect analysis (G1.2)	VR1.2.2	Label sensitivity analysis
Inaccurate data labels	Data collection & preparation	Automatic detection of inaccurate labels (G1.3)	VR1.3	Confident Learning metric
Problems with synthetic data	Data collection & preparation	Reality gap in the model's performance (G2.1)	VR2.1	Reality gap performance test
Problems with synthetic data	Data collection & preparation	Reality gap in the model's perception (G2.2)	VR2.2	Neural activation metric
Lack of robustness	Modeling	Lack of adversarial robustness (G3.1)	VR3.1	Argumentation
Lack of robustness	Modeling	Lack of robustness against natural variation in the input domain (G3.2)	VR3.2	Perturbation-based robustness quantification
Lack of robustness	Modeling	Lack of robustness against counterfactuals (G3.3)	VR3.3	Counterfactual analysis

Abbildung 26: 2 Wege zur Darstellung der Landscape of AI Safety Concerns

Die GSN-basierte visuelle Darstellung bietet eine strukturierte, klare und prägnante Visualisierung der AI Safety Concerns, Mitigations-Maßnahmen und Nachweisen. Sie ist

besonders geeignet, um sich schnell einen Überblick zu verschaffen und erleichtert die klare Kommunikation komplexer Sicherheitsargumente und -strategien.

Die tabellarische Darstellung hingegen bietet einen umfassenden und detaillierten Überblick, indem sie jedes AI Safety Concerns neben expliziten Mitigations-Maßnahmen, Validierungsmetriken und Lebenszyklusphasen auflistet. Diese detaillierte Tabellierung ermöglicht analytische Einblicke und unterstützt detaillierte Audits und präzise Bewertungen.

2.1.2.2 Metriken zur Bewertung der Vertrauenswürdigkeit von KI-Funktionen

In Zusammenarbeit von vielen Konsortialpartnern konnten im Projekt 20 Metriken identifiziert und implementiert werden, welche sich in die folgenden 4 Themen kategorisieren lassen:

- Modell Performanz
    - Fusion for Object Detection - Cycle time
    - Fusion for Object Detection - Number of detected objects
    - Library of metrics for Sensor Fusion
    - QI² network analysis
    - Per-Class IoU
    - Prototype-based zero-shot OOD detection and segmentation
    - Metrics for OoD Detection

- Robustheit
    - Data-Driven Verification
    - Semantic Performance Discrepancy
    - Robustness Quantification
    - Robustness Verification

- Erklärbarkeit & Transparenz
    - Metrik “ECS - Equivalence class Sets” Metric “QI² - Network Analysis”
    - Metrik “ODD concept coverage metric”
    - Metrik “Bias & Fairness”
    - Method “Visual exploration of GT image semantics”
    - Method “Visual Inspection Coverage”
    - Metrik “Saliency Map Metric”

- Datenqualität
    - K-Projection Coverage
    - QI2 Data Quality
    - Regional Max SHLQI2

Neben den oben aufgeführten State-of-the-Art-Analysen wurden sogenannte „Fact Sheets“ je Metrik erstellt, in denen in standardisierter Weise einen Überblick über die Eigenschaften, Funktionalitäten (in Bezug auf den Sicherheitsnachweis) und Limitationen der einzelnen Metriken gegeben wird. Dabei wird auch Bezug zu den AI Safety Concerns aus der Landscape of AI Safety Concerns herstellt. Die Metriken und deren Dokumentation bilden einen wichtigen Baustein für den Sicherheitsnachweis.

Metriken
Metrics for OoD Detection
Prototype-based zero-shot OOD detection and segmentation
Per-Class IoU
Bias & Fairness
ODD Concept Coverage
ECS
QI² network analysis
Saliency map metric
Visual Inspection Coverage
Visual exploration of GT image semantics
Data-driven Verification
Robustness Certification
Robustness Quantification
Semantic Performance Discrepancy
Fusion for Object Detection- Cycle time
Fusion for object detection- Number of detected objects
Library of metrics for Sensor Fusion
K-Projection Coverage
QI² Dataquality
Regional Max SHLQI2

Tabelle 1: Übersicht Metriken

Siemens Mobility konnte sich hier insbesondere bei folgenden Metriken intensiv einbringen:

Saliency Maps

Dies sind entscheidende Werkzeuge zur Verbesserung der Erklärbarkeit von KI-basierten Perzeptionssystemen, insbesondere in sicherheitskritischen Anwendungen wie dem vollautomatisierten Fahren von Regionalzügen. Eine Saliency Map stellt eine Art „Karte“ dar, welche visuell zeigt, welche Bereiche innerhalb der Eingabedaten die Vorhersagen des KI-Modells signifikant beeinflussen. Sie helfen damit den Stakeholdern, den Entscheidungsprozess des Modells besser zu verstehen.

Die Methodik zur Erstellung von Saliency Maps umfasst die Berechnung der Ausprägung von Modellvorhersagen, die Schwellenwertbildung der Ausprägungswerte und die Durchführung einer Intersection over Union (IoU)-Analyse zwischen der Saliency Map und den Ground-Truth-Daten. Diese Quantifizierung und Visualisierung der Fokusbereiche des Modells liefert umsetzbare Erkenntnisse über das Verhalten des Modells und ermöglicht gezielte Verbesserungen und Validierungen.

Saliency Maps dienen als wesentliche Grundlage für die weitere Erforschung und Entwicklung von Erklärbarkeitsmethoden und tragen erheblich zur Robustheit und Transparenz von KI-Modellen bei. Durch die klare Hervorhebung der Entscheidungsbereiche spielen sie eine entscheidende Rolle in der Sicherheitsargumentation und der Einhaltung von Vorschriften, unterstützen eine rigorose Sicherheitsvalidierung und fördern das Vertrauen der Stakeholder.

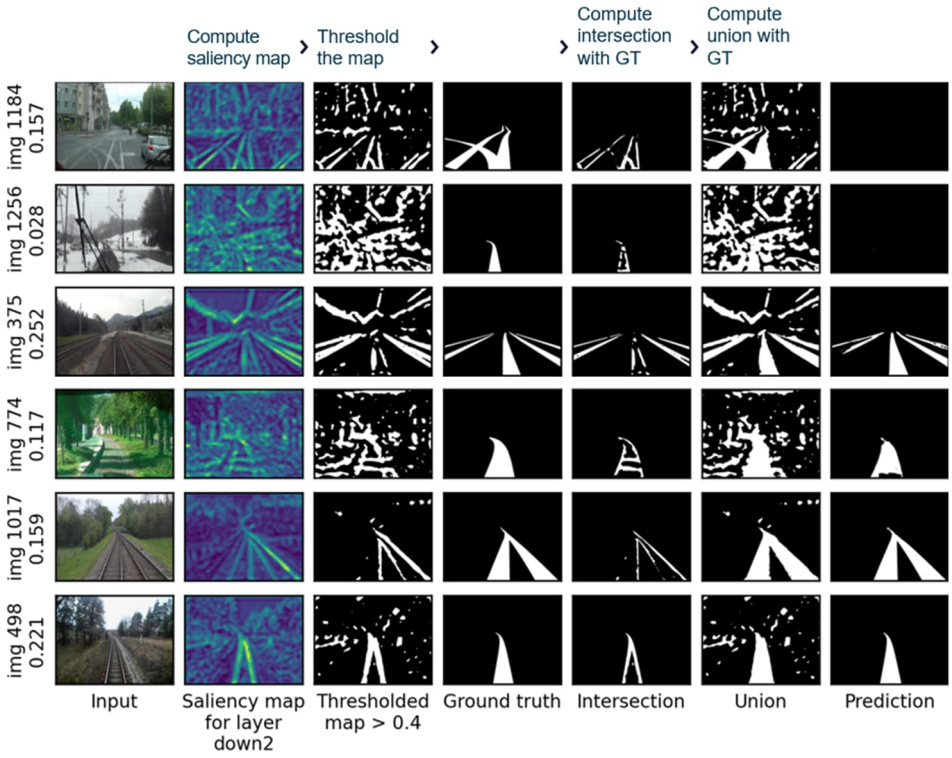


Abbildung 27: Saliency Map Beispiele

QI² (Qualitätsindikator)

Metrik zur quantitativen Bewertung von ML-Datensätzen. Mit dieser können Schwachstellen im Datensatz gezielt aufgespürt werden. QI² bietet einen neuartigen Ansatz, um die komplexen Beziehungen innerhalb von Datensätzen zu visualisieren und zu analysieren, was für die Entwicklung robuster und zuverlässiger KI-Systeme von entscheidender Bedeutung ist.

Im Kern ermöglicht QI² eine komprimierte und repräsentative Visualisierung der lokalen Input-Output-Beziehungen innerhalb eines Datensatzes. Dies bedeutet, dass die Methode nicht nur die globalen Eigenschaften der Daten betrachtet, sondern auch, wie sich einzelne Datenpunkte oder kleine Gruppen von Datenpunkten zueinander verhalten und welche Auswirkungen sie auf die Ausgabe eines Modells haben könnten *[Sicherichs, C.; Geerkens S.; Braun, A.; Waschulzik, T. (2022) "QI2 - an Interactive Tool for Data Quality Assurance"].*

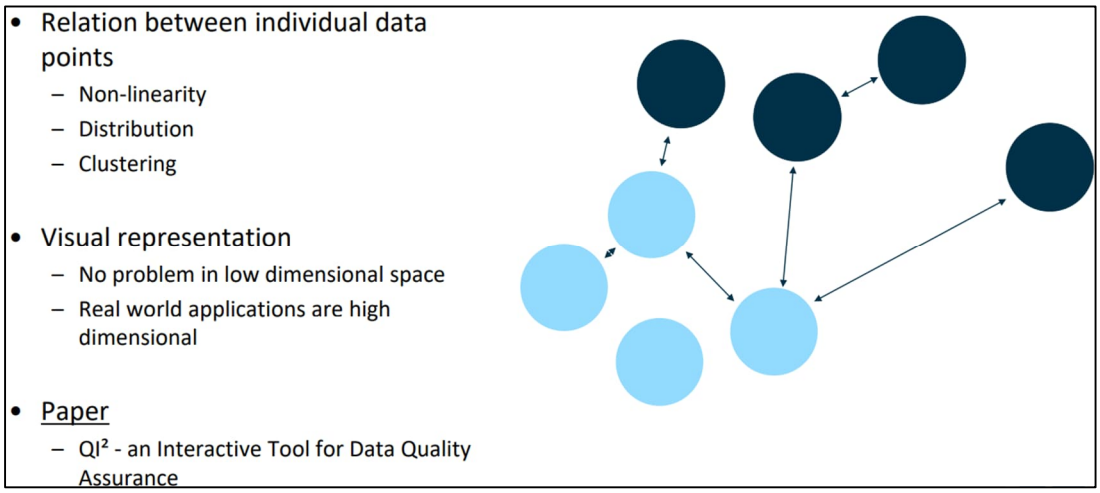


Abbildung 28: QI² - integrierter Qualitätsindikator (I)

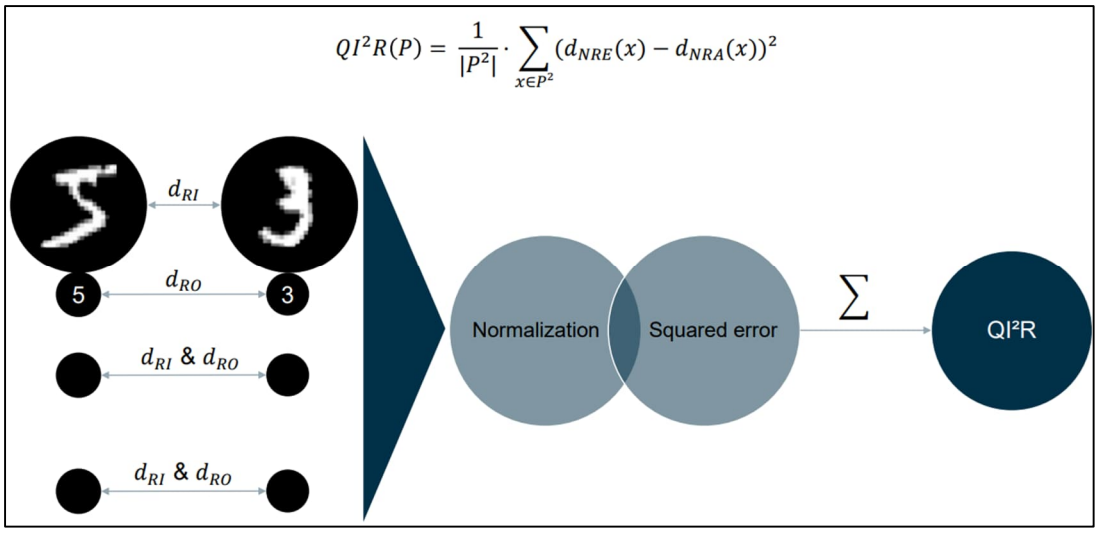


Abbildung 29: QI² - integrierter Qualitätsindikator (II)

Ein zentrales Element von QI² ist die sogenannte SHLQI²-Visualisierung (Shape-based Local QI²). Diese Visualisierung erlaubt es, "interessante" Datenpunkte und spezifische Qualitätsaspekte schnell zu identifizieren. Dazu gehören:

- Ausreißer (Outliers):
  Datenpunkte, die sich signifikant von der Mehrheit der Daten unterscheiden und oft auf Messfehler oder ungewöhnliche Ereignisse hinweisen.

- Lineare (einfache) Unteraufgaben (Linear Subtasks):
  Bereiche im Datensatz, in denen die Beziehungen zwischen Input und Output linar (einfach) sind.

- Diskontinuitäten (Discontinuities):
  Plötzliche Sprünge oder Brüche in den Datenbeziehungen, die auf Schwellenwerte, Fehler oder wichtige Übergänge hinweisen können.

QI² kann sowohl für die globale als auch für die lokale Datenqualitätsanalyse eingesetzt werden. Bei der globalen Analyse wird die Gesamtkomplexität des Datensatzes bewertet, während die lokale Analyse mithilfe der SHLQI²-Visualisierung spezifische Qualitätsprobleme aufdeckt.

Die Funktionsweise unterscheidet sich leicht je nach Art der Aufgabe:

- Für Approximationsaufgaben (z.B. Regression):
  Das SHLQI²-Histogramm liefert Einblicke in die Komplexität der Input-Output-Beziehungen. Höhere Werte im Histogramm weisen auf komplexere Beziehungen hin, verglichen mit einem zufällig verteilten Datensatz. Dies hilft zu verstehen, wie "schwierig" bestimmte Bereiche des Datensatzes für ein Approximationsmodell sind.

- Für Klassifikationsaufgaben:
  SHLQI² zeigt charakteristische steile Anstiege und plötzliche Abfälle in der Komplexität. Diese Muster korrespondieren direkt mit Klassengrenzen und homogenen Clustern von Datenpunkten. Das ermöglicht es, die Trennbarkeit von Klassen zu beurteilen und Bereiche mit klar definierten oder unscharfen Grenzen zu identifizieren.

Die Anwendbarkeit von QI² wurde am Beispiel des MNIST-Datensatzes (handgeschriebene Ziffern) demonstriert. Hier konnte die Methode erfolgreich homogene Cluster von Ziffern, Datenpunkte, die außerhalb der erwarteten Verteilung lagen (Out-of-Distribution-Daten), sowie einfache Untergruppen innerhalb der Daten identifizieren, die potenziell separat behandelt werden könnten.

Der Hauptnutzen von QI² liegt in seiner Fähigkeit, eine umfassende und interaktive Datenqualitätssicherung zu ermöglichen. Im Gegensatz zu traditionellen Methoden, die oft nur globale Statistiken liefern, erlaubt QI² einen tiefen Einblick in die Mikrostruktur der Daten. Die Vorteile für die Datenqualitätssicherstellung sind vielfältig:

- Früherkennung von Problemen:
  Durch die Visualisierung und Analyse lokaler Beziehungen können Datenqualitätsprobleme wie Ausreißer, fehlerhafte Messungen oder inkonsistente Daten frühzeitig im Entwicklungsprozess von KI-Systemen erkannt werden.

- Verbesserte Modellleistung:
  Das Verständnis der Datenqualität hilft dabei, Trainingsdaten zu bereinigen und zu optimieren. Dies führt zu robusteren und genaueren KI-Modellen, da das Modell auf einer fundierteren quaslitätsgesicherten Datenbasis trainiert wird.

- Effiziente Datenbereinigung:
  Die Identifizierung von linearen Unteraufgaben oder homogenen Clustern kann die Datenbereinigung und -vorverarbeitung effizienter gestalten, da man sich auf die komplexen oder problematischen Bereiche konzentrieren kann.

- Umgang mit komplexen Datensätzen:
  Besonders bei großen und komplexen Datensätzen, bei denen manuelle Prüfungen kaum möglich sind, bietet QI² eine skalierbare Lösung zur Identifizierung von Qualitätsproblemen und gibt Hinweise, ob bestimme Vorverarbeitungsschritte in der Lage sind die globale und oder lokale Komplexität zu reduzieren.

- Unterstützung der KI-Entwicklung:
  QI² dient als wertvolles Werkzeug für KI-Entwickler, um die Qualität ihrer Trainingsdaten zu beurteilen, Modellfehler auf Datenprobleme zurückzuführen und somit den gesamten Entwicklungsprozess zu beschleunigen und zu verbessern.

- Erkennung von Out-of-Distribution-Daten**:**
  Die Fähigkeit, Datenpunkte außerhalb der erwarteten Verteilung zu erkennen, ist entscheidend für die Sicherheit und Zuverlässigkeit von KI-Systemen, insbesondere in kritischen Anwendungen.

Zusammenfassend lässt sich sagen, dass QI² ein leistungsstarkes neues Werkzeug für die Datenwissenschaft und KI-Entwicklung darstellt. Es ermöglicht eine detaillierte und interaktive Analyse der Datenqualität, die weit über herkömmliche Methoden hinausgeht und somit maßgeblich zur Entwicklung hochwertiger und vertrauenswürdiger KI-Systeme beiträgt.

ECS (Equivalent Class Sets)

Die Metrik ECS ist ein neuartiger Ansatz zur Datenqualitätssicherung. Sie zielt darauf ab, versteckte Ungleichgewichte oder fehlende Varianz in Datensätzen zu identifizieren und so die Datenqualität umfassend zu bewerten.

Die ECS-Metrik basiert auf dem Umstand, dass bei Aufgaben mit überwachtem Lernen ein Datensatz in Eingabedaten (Features, die zur Vorhersage dienen) und Ausgabedaten (die vorhergesagten Werte) unterteilt werden kann. Um den ECS zu berechnen, werden zwei Metriken benötigt: eine für den Abstand im Eingaberaum (Input Space) und eine für den Abstand im Ausgaberaum (Output Space). Alle Feature-Werte müssen numerisch sein; nicht-numerische Daten müssen entsprechend umgewandelt werden [*Sicherichs, C.; Geerkens S.; Braun, A.; Waschulzik, T. (2022) "ECS - an Interactive Tool for Data Quality Assurance"*].

Die Kernidee des ECS ist der Vergleich von Datenpunktpaaren. Für jedes Paar von Datenpunkten werden die Abstände im Eingabe- und Ausgaberaum analysiert. Dabei ergeben sich vier grundlegende Szenarien:

- input space / output space
- equivalent / unequivalent

		Output space	
		Equivalent	Unequivalent
Input space	Equivalent	EE	EU
	Unequivalent	UE	UU

Abbildung 30: Legende zur Erklärung der Äquivalenz

- Kleiner Abstand im Eingaberaum– Kleiner Abstand im Ausgaberaum (ECS_EE): Datenpunkte, die sich im Eingaberaum ähneln und auch im Ausgaberaum ähnliche Ergebnisse liefern. Dies deutet oft auf typische Anwendungsfälle hin.

- Kleiner Abstand im Eingaberaum – Großer Abstand im Ausgaberaum (ECS_EU): Datenpunkte, die im Eingaberaum ähnlich sind, aber im Ausgaberaum stark unterschiedliche Ergebnisse zeigen. Dies kann auf problematische Bereiche hinweisen, z.B. wenn eine kleine Änderung im Input zu einem großen Sprung im Output führt.

- Großer Abstand im Eingaberaum – Kleiner Abstand im Ausgaberaum (ECS_UE): Datenpunkte, die im Eingaberaum weit voneinander entfernt sind, aber dennoch ähnliche Ausgaben haben. Dies könnte auf Redundanzen oder unerwartete Korrelationen hindeuten.

- Großer Abstand im Eingaberaum – Großer Abstand im Ausgaberaum (ECS_UU): Datenpunkte, die sich sowohl im Eingabe- als auch im Ausgaberaum stark unterscheiden. Dies sind oft unkritische Vergleiche.

Diese Vergleiche werden in sogenannten "ECS-Sets" gespeichert. Jedes der vier ECS-Sets repräsentiert alle Datenpunktvergleiche, die in eines der vier Szenarien fallen. Durch die Analyse dieser Sets, oft visualisiert als Histogramme oder Funktionen, kann das ECS Muster und Auffälligkeiten in den Daten erkennen. Die Visualisierung zeigt, wie viele Datenpunktpaare die jeweiligen Abstandsbeziehungen aufweisen. Steile und frühe Anstiege in diesen Funktionen weisen auf das Vorhandensein von vielen Datenpunkten mit kleinen Eingabe- und Ausgabeabständen hin.

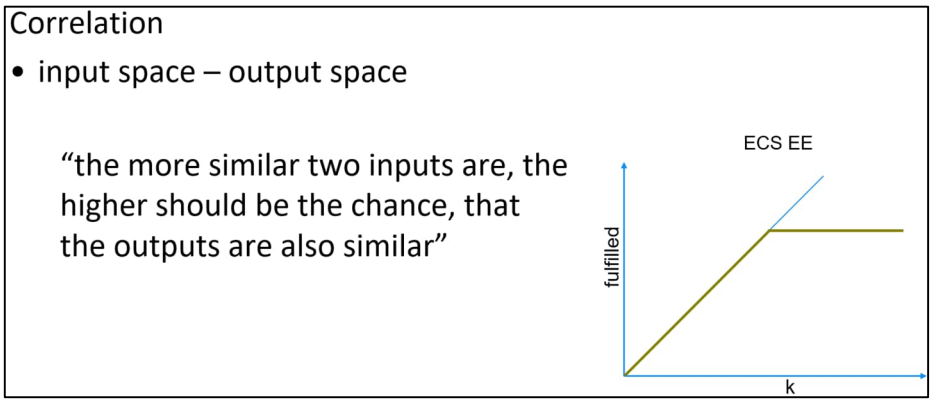


Abbildung 31: ECS - grafische Darstellung eines idealen Datensatzes

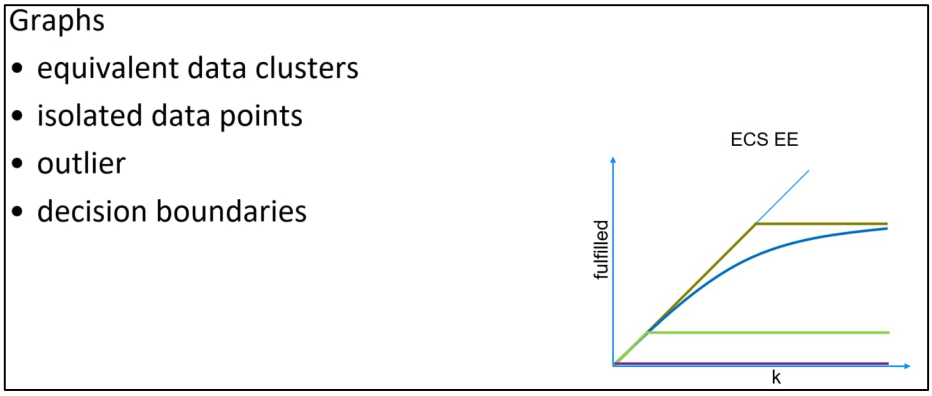


Abbildung 32: ECS - grafische Darstellung eines realen Datensatzes

Das ECS ist vielseitig einsetzbar, um verschiedene Datenqualitätseigenschaften zu erkennen. Typische Anwendungsszenarien umfassen:

- Erkennung von Ausreißern (Outlier Detection):
  Das ECS_EU (kleiner Abstand im Eingaberaum, großer im Ausgaberaum) ist besonders nützlich, um Ausreißer zu identifizieren. Ein Datenpunkt, der trotz ähnlicher Nachbarn im Eingaberaum eine stark abweichende Ausgabe hat, wird als potenzieller Ausreißer markiert.

- Identifizierung isolierter Datenpunkte:
  Sowohl ECS_UE als auch ECS_UU können verwendet werden, um Datenpunkte zu finden, die nur wenige Nachbarn in ihrer Nähe haben, was auf isolierte oder seltene Fälle hindeutet.

- Auffinden lokaler Gruppen mit identischer Ausgabe:
  Das ECS_EE (kleiner Eingabeabstand, kleiner Ausgabeabstand) eignet sich hervorragend, um Gruppen von Datenpunkten zu erkennen, die im Eingaberaum eng beieinander liegen und dieselbe Ausgabe produzieren. Dies ist wichtig, um die Konsistenz von Labels oder Klassifikationen zu prüfen.

- Validierung quantitativer Datenanforderungen:
  Das ECS kann verwendet werden, um spezifische Anforderungen an die Datenqualität zu überprüfen, z.B. die Mindestanzahl von Elementen pro Gruppe, die maximale Anzahl von Ausreißern oder die Anzahl lokaler Gruppen.

- Korrelation von Eingabe- und Ausgaberaum:
  Es ermöglicht die Identifizierung von Bereichen im Eingaberaum, die mit bestimmten Ausgaben im Ausgaberaum korrelieren.

Der Hauptnutzen des ECS liegt in seiner Fähigkeit, die Datenqualität umfassend und effizient zu sichern, insbesondere für ML-Systeme, die in kritischen Bereichen eingesetzt werden:

- Umfassende Datenqualitätsbewertung:
  Im Gegensatz zu vielen anderen Methoden, die sich oft auf eine spezifische Datenqualitätseigenschaft konzentrieren, kann das ECS mit einem einzigen Ansatz mehrere Eigenschaften gleichzeitig analysieren. Dies spart Zeit und Ressourcen.

- Erkennung schädlicher Datenpunkte:
  Das ECS ist besonders geeignet, Datenpunkte zu erkennen, die potenziell schädliche Eigenschaften für die Modellbildung oder den Betrieb von ML-Systemen aufweisen könnten. Dies ist entscheidend, um die Robustheit und Zuverlässigkeit von ML-Modellen zu gewährleisten.

- Arbeit mit Originaldaten:
  Ein großer Vorteil des ECS ist, dass es auf den Originalwerten der Daten operiert. Viele andere Methoden zur Datenanalyse erfordern eine Dimensionsreduktion, die oft mit einem erheblichen Informationsverlust einhergeht. Das ECS vermeidet diesen Verlust und berücksichtigt alle verfügbaren Informationen.

- Interaktive Analyse:
  Das ECS bietet eine interaktive Möglichkeit zur Analyse von Daten. Auch wenn die Roh-ECS-Sets für Menschen schwer lesbar sind, ermöglichen die daraus abgeleiteten Visualisierungen (Histogramme, Funktionen) eine intuitive Interpretation der Datenqualität.

- Anpassungsfähigkeit:
  Durch die Wahl geeigneter Metriken und Schwellenwerte kann das ECS an verschiedene Datentypen und spezifische Anforderungen angepasst werden, selbst bei Daten mit vielen Features und dem Problem des "Fluchs der Dimensionalität".

- Unterstützung für sicherheitskritische Systeme:
  Für Siemens ist das ECS besonders relevant, da es die Grundlage für eine zuverlässige Datenbasis in sicherheitskritischen Anwendungen (z.B. im Bahnverkehr oder in der Automatisierung) schafft, wo Datenfehler schwerwiegende Folgen haben können. Es hilft, die Qualität von Trainingsdaten für ML-Modelle sicherzustellen, was für die Validierung und Zertifizierung solcher Systeme unerlässlich ist.

Zusammenfassend bietet das ECS einen leistungsstarken und flexiblen Rahmen zur systematischen Analyse und Sicherstellung der Datenqualität, indem es auf der lokalen Ähnlichkeit von Datenpunkten basiert und die komplexen Beziehungen zwischen Eingabe- und Ausgabedaten transparent macht.

Beide vorgenannten Metriken (ECS und QI²) wurden gemeinsam mit dem Projektpartner Hochschule Düsseldorf entwickelt.

Eine weitere Metrik, die durch Siemens Mobility konzipiert wurde, ist der Regional Max SHLQI², welcher auf der SHLQI²-Visualisierung des QI² aufbaut.

Der SHLQI² visualisiert die lokale Komplexitätsverteilung eines Datensatzes. Der RM SHLQI² fasst diese komplexe Struktur zu einer kompakten Kurve zusammen und ermöglicht so, Unterschiede zwischen Datensätzen sichtbar und quantifizierbar zu machen, z.B. zur Erkennung von Fehlern oder zur Bewertung der Ähnlichkeit zwischen realen und simulierten Daten.

Die Metriken ECS, QI² und Regional Max SHLQI² sind Teil der sogenannten QUEEN-Methodik (Quality-aware Efficient Evolution of Neural networks), die darauf abzielt, einen effizienteren KI-Entwicklungsprozess mit eingebauter Qualitätssicherung zu schaffen.

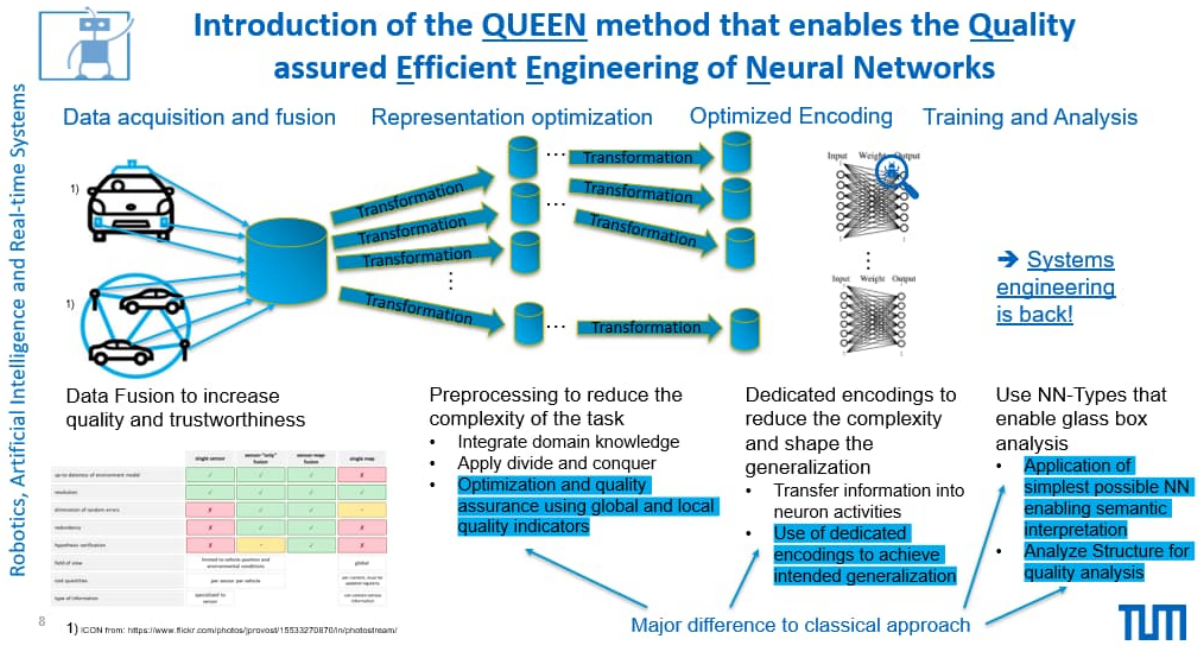


Abbildung 33: Darstellung der QUEEN Methodik – Die QUEEN-Methodik ist inzwischen auch Lehrinhalt der Vorlesungen „Autonomes Fahren“ und „Vertrauenswürdige Systeme mit Maschinellen Lernen“ an der Technischen Universität München

Neben der Identifikation der Metriken wurde sich in diesem Arbeitspaket auch mit der Verifikation von KI-Funktionen an sich beschäftigt. Hierfür wurden Bilddaten hinsichtlich globaler Veränderungen (z. B. Wetter, Rauschen, Helligkeit) sowie gegen Veränderungen und Variation auf Objekt-Level synthetisch erzeugt, um die KI-Funktionen gegen diese Einflüsse evaluieren zu können.

Ebenso war die Entwicklung einer Methodik zur Beschreibung der Betriebsumgebung in diesem Arbeitspaket geplant. Das so entstandene methodische Framework zur Beschreibung des Umgebungskontexts, in welchem das KI-System eingebettet ist, galt als Basis für die weiteren Arbeiten an der Operation Design Domain (siehe Kapitel 2.1.3.2).

2.1.3 AP3: Fahrzeugarchitektur im GoA4-Betrieb mit dem Fokus auf sichere KI-basierte Funktionen

Dieses Arbeitspaket wurde von Siemens Mobility geleitet. Sie übernahm die Koordination der inhaltlichen Arbeiten sowie die Abstimmung zwischen den verschiedenen Unterarbeitspaketen und anderen Arbeitspaketen des Projekts, insbesondere AP2 und AP4.

In diesen Arbeitspaket wurden zunächst die (Sicherheits-) Anforderungen an ein KI-basiertes Perzeptionssystem und die Betriebsumgebung (Operational Design Domain = ODD) definiert, um darauf aufbauend eine Architektur für ein solches Systems ableiten zu können. Diese Architektur inkl. Sicherheits-Muster (wie z. B. Redundanzen) bildete wiederum den Input für die Implementierung des prototypischen Testsystems (System under Tests = SuT). Auf Basis der erstellten Architektur und der Implementierung des System under Tests sollte eine Sicherheitsargumentation für ein KI-basiertes Perzeptionssystem aufgestellt werden.

2.1.3.1 Anforderungen

Die Anforderungen an ein KI-basiertes Perzeptionssystem umfassen neben der Umgebungswahrnehmung, der Erkennung von Hindernissen und den daraufhin auszulösenden Reaktionen ebenso Anforderungen an die Integration und Wartung eines solchen KI-basierten Systems.

Die zentralen sicherheitsrelevanten Funktionen, welche im wesentlichen KI-basiert umgesetzt wurden, sind die „Erkennung von Personen“ („Person on track“, siehe Abbildung 34) und die „Erkennung von (für ein Schienenfahrzeug) gefährlichen Gegenständen“ („Passenger in train“ siehe Abbildung 34) auf oder neben dem eigenen Fahrweg. Dies entspricht den aktuell für einen Triebfahrzeugführer:innen (Tf) gültigen Regularien (wie z. B. DB RIL 408.2341). Diese besagt: „Der Tf des Fahrzeuges, das an der Spitze eines Zuges fährt, muss die zu befahrende Strecke, die Signale, die Bahnübergänge und die Oberleitung beobachten. Dabei muss er auf Unregelmäßigkeiten achten, die den Zug gefährden könnten.“ [*Deutsche Bahn AG, "Richtlinie 408.2341: Anforderungen an Triebfahrzeugführer hinsichtlich Streckenbeobachtung"*]

Für diese sicherheitsrelevanten Funktionen wurden betriebliche Anwendungsfälle (Use Cases) als Grundlage für alle weiteren Betrachtungen im Projekt erstellt.

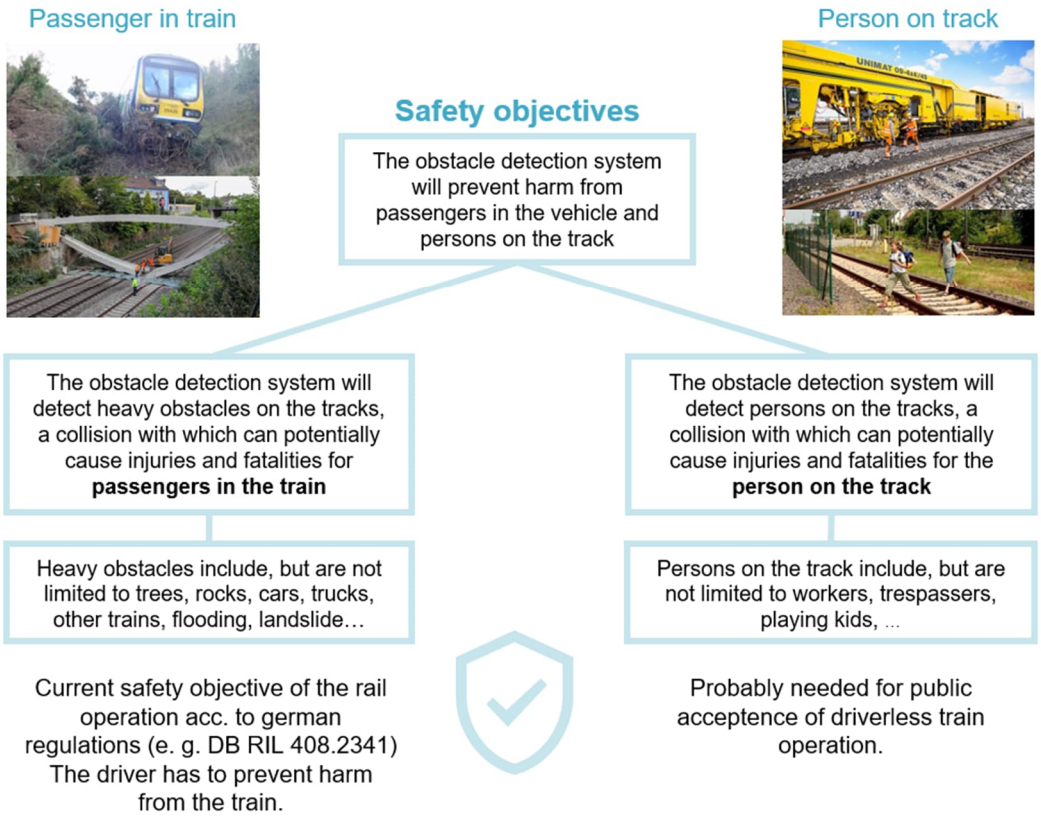


Abbildung 34: Darstellung der beiden Use Cases

Die Leistungsfähigkeit eines Triebfahrzeugführer wird nach Hinzen [*Hinzen, A. 1993, Der Einfluss des menschlichen Fehlers auf die Sicherheit der Eisenbahn, Aachen, Rhein.-Westfäl. Technische Hochschule Aachen*] mit einer Fehlerwahrscheinlichkeit von 1% bewertet.

Abweichend vom geplanten Ansatz der Erstellung eines technischen Sicherheitsplan (TeSip), in dem über eine explizite Risikoanalyse die Gefährdungen und Risiken ermittelt werden, wurde im Projekt festgelegt, die Ableitung der Gefährdungen und Risiken über den Vergleich zu dem Referenzsystem „Triebfahrzeugführer“ gemäß CSM RA (Common Safety Method for Risk Evaluation and Assessment) zu ermitteln. Es wurde entschieden, den Sicherheitsnachweis gemäß der Norm EN 61508 durchzuführen.

Die in AP1 abgeleitete PFD von 1% bildet die grundlegende Sicherheitsanforderung an eine voll- automatische Fahrwegsüberwachung. Darüber hinaus wurde durch Siemens Mobility im Projekt ein ganzes Set an weiteren Anforderungen an ein solches System identifiziert und definiert. Dabei bestand die Herausforderungen darin, die quantitativen Anforderungen an einen menschlichen Triebfahrzeugfahrer in messbare und testbare Anforderungen für ein KI-basiertes System zu überführen. Der aktuelle Zugbetrieb mit einem menschlichen Tf ist durch Vorschriften geregelt, welche auf dem qualitativen Urteilsvermögen und dem gesunden

Menschenverstand basieren. Zum Beispiel verstehen menschliche Fahrer intuitiv Konzepte wie "sichere Abstände" oder "Unregelmäßigkeiten auf der Strecke", während für ein KI-basiertes System diese Parameter explizit quantifizieren werden müssen. Dies macht eine direkte Überführung in Anforderungen für ein KI-basiertes System herausfordernd.

Im Projekt Anforderungen u.a. hinsichtlich folgender Aspekte definiert:

- In welcher Entfernung muss welche Objektgröße bei welcher Helligkeit und Reflektivität erkannt werden?
- Wie könne Objekte in unterschiedliche Klassen unterteilt werden?
    - Menschen
    - Tiere
    - Andere Verkehrsteilnehmer
    - Landmarken, statische Objekte
    - etc.
- Für welche Objekte muss eine Reaktion eingeleitet werden?
    - Black list (vom Betreiber festzulegen)
- Für welche Objekte muss keine Reaktion eingeleitet werden?
    - White list (vom Betreiber festzulegen)
- In welche Zonen kann der Bereich vor dem Fahrzeug eingeteilt werden?

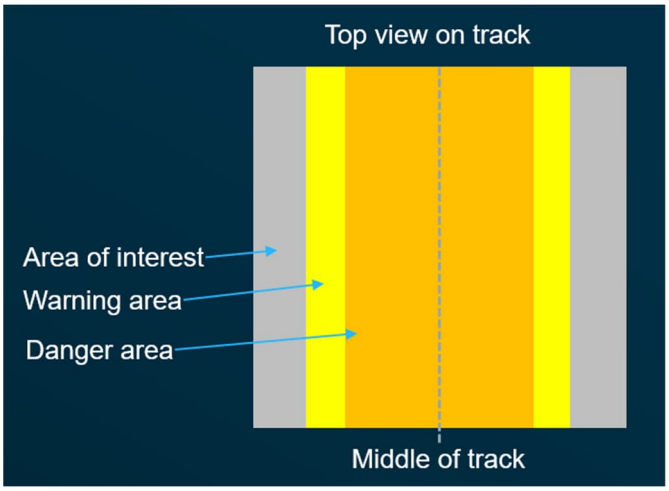


Abbildung 35: Definition der Bereiche vor dem Fahrzeug

Ein KI-basiertes Perzeptionssystem muss räumliche und betriebliche Grenzen explizit definieren, um Umweltbeobachtungen effektiv in umsetzbare Sicherheitsinformationen zu kategorisieren. Das safe.trAIn-Projekt adressiert dieses Bedürfnis durch die klare Definition von drei kritischen Zonen. Jede Zone hat spezifische Anforderungen basierend auf dem Risikoniveau, das sie darstellt:

- Beobachtungsbereich
  Der Beobachtungsbereich umfasst die größte Zone um den Zug bzw. vor der Zugfront. Innerhalb dieser Zone beobachtet das Perzeptionssystem kontinuierlich, aber Objekte darin führen nicht direkt zu einer Gefährdung. Objekte werden identifiziert und überwacht, aber keine sofortige Reaktion ist notwendig.

- Warnbereich
  Der Warnbereich, der sich näher an dem Fahrschlauch (Lichtraumprofil) des Zuges befindet, ist für die Früherkennung potenzieller Bedrohungen vorgesehen. Objekte, die in diesen Bereich auftauchen, veranlassen das System, sich auf mögliche Vorsichtsmaßnahmen oder die Bereitschaft zum Übergang zu Notfallverfahren vorzubereiten.

- Gefahrenbereich
  Der Gefahrenbereich befindet sich am nächsten zu oder überlappt mit dem Lichtraumprofil des Zuges. Jedes Objekt, das in diese Zone eintritt, löst eine sofortige Sicherheitsreaktion aus, um eine Kollision mit einem Objekt zu verhindern.

Diese klaren, quantitativen Abgrenzungen ermöglichen es einem KI-basierten Perzeptionssystem, schnell und angemessen zu handeln.

Bei der Definition der Anforderungen wurde auch festgelegt, ob die jeweilige Anforderung eine Sicherheitsrelevanz besitzt oder nicht.

2.1.3.2 Operational Design Domain

Ein zentrales Element für die Entwicklung von KI-basierten automatisierten Systemen ist die Operational Design Domain (= ODD). Sie beschreibt die Einsatzbedingungen, unter denen das System funktionieren soll, inkl. der Umgebungs-/Wetter-Bedingungen, geografischer Bedingungen, Lichtverhältnisse, etc. Zur ODD zählen auch andere Verkehrsteilnehmer oder statische Elemente wie Elemente der Infrastruktur (z. B. Landmarken).

Die ODD fungiert als grundlegendes Element bei der Entwicklung und Implementierung KI-basierter Perzeptionssysteme. Ihre Rolle geht über die bloße Definition von Betriebsbedingungen hinaus. Die ODD hat Einfluss auf den gesamten System-/KI-Entwicklungsprozess und bietet wesentliche Leitlinien für jede Phase des Entwicklungszyklus

- auf die Anforderungen an die Architektur und die Implementierung,
- auf den Sicherheitsnachweis,
- auf das Training des Systems (Trainingsdaten),
- auf die Validierung (Testdaten, Testabdeckung),
- auf das Monitoring während des Betriebes (zur Erkennung, ob das System den vorgesehenen Einsatzbereich verlässt – Out of Distribution = OoD).

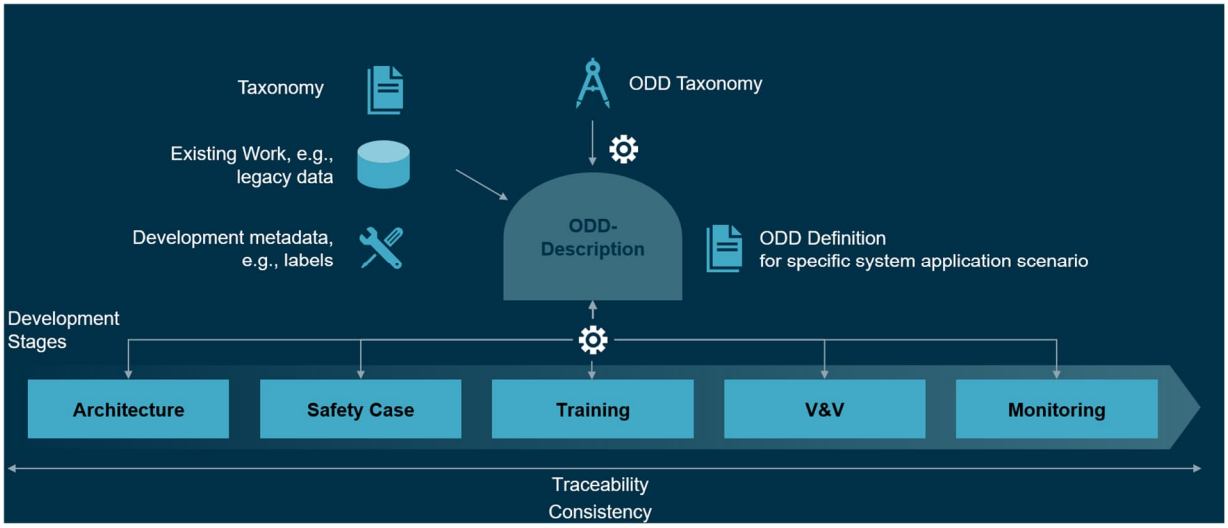


Abbildung 36: Operational Design Domain als zentrales Element im Entwicklungsprozess

Da die Granularität der von der ODD benötigten Informationen in den unterschiedlichen Entwicklungsstufen variiert, wurde im Projekt ein entsprechender Prozess aufgesetzt und die nötigen Werkzeuge entwickelt, so dass die ODD von allen Stakeholdern im KI-Entwicklungsprozess entsprechend gut genutzt und im jeweils benötigten Format integriert werden kann. Damit sind verschiedene Exporte verfügbar, wie z. B. als Python Integration, Eclipse oder ein Visualisierungstool.

Durch die klare Definition der Betriebsgrenzen und -bedingungen stellt die ODD sicher, dass jede Entwicklungsaktivität konsequent auf realistische und relevante Szenarien abgestimmt ist. Diese Abstimmung verbessert erheblich die Qualität und Effektivität der Systemvalidierung und -verifikation und stellt sicher, dass KI-basiertes Perzeptionssystem robust, zuverlässig und in der Lage ist, alle festgelegten Betriebsbedingungen sicher zu bewältigen. Darüber hinaus verbessert die zentrale Rolle des ODD die Rückverfolgbarkeit, indem Entwickler und Sicherheitsprüfer in die Lage versetzt werden, die Systemleistung im Vergleich zu klar dokumentierten Betriebsszenarien umfassend zu verfolgen und zu bewerten.

Die explizite Definition dieser Umgebungsbedingungen gewährleistet Klarheit hinsichtlich der Systemfähigkeiten und -einschränkungen und erhöht die Sicherheit eines KI-basierten Perzeptionssystems. Die ODD gibt zulässige Szenarien (z.B. klares Wetter, Tageslicht) und

eingeschränkte Bedingungen (z.B. extremes Wetter, schlechte Sicht) an, unter denen ein vollautomatischer Zug entweder die Betriebsgeschwindigkeit reduzieren muss, zusätzliche menschliche Überwachung erforderlich ist oder der vollautomatische Betrieb vollständig eingestellt werden müssen.

Im Projekt safe.trAIn wurde in enger Zusammenarbeit von Siemens Mobility und Fraunhofer IKS erstmalig eine ODD mit einer entsprechenden Taxonomie für den Bahnsektor entwickelt. Dabei wurde sich an anderen Domänen wie z. B. Automotive (PAS 1883 oder ASAM OpenODD) orientiert. Diese mussten aber an vielen Stellen um die entsprechenden Bahn-Spezifika erweitert und angepasst werden.

Die Taxonomie teilt die Betriebsbedingungen in strukturierte, klar definierte Kategorien ein. Dies ist entscheidend für die einheitliche Beschreibung von Umwelt-, geografischer und zeitlicher Einschränkungen, welche ein KI-basiertes System einhalten muss. Durch die Nutzung einer gut strukturierten Taxonomie können Ingenieure und Sicherheitsprüfer die genauen Bedingungen, unter denen KI-Systeme funktionieren sollen, einheitlich dokumentieren, bewerten und kommunizieren. Diese systematische Klassifizierung hilft, Mehrdeutigkeiten und Inkonsistenzen zu vermeiden und sorgt für eine umfassende und klare Dokumentation, die für Systemtests und Validierungen erforderlich ist.

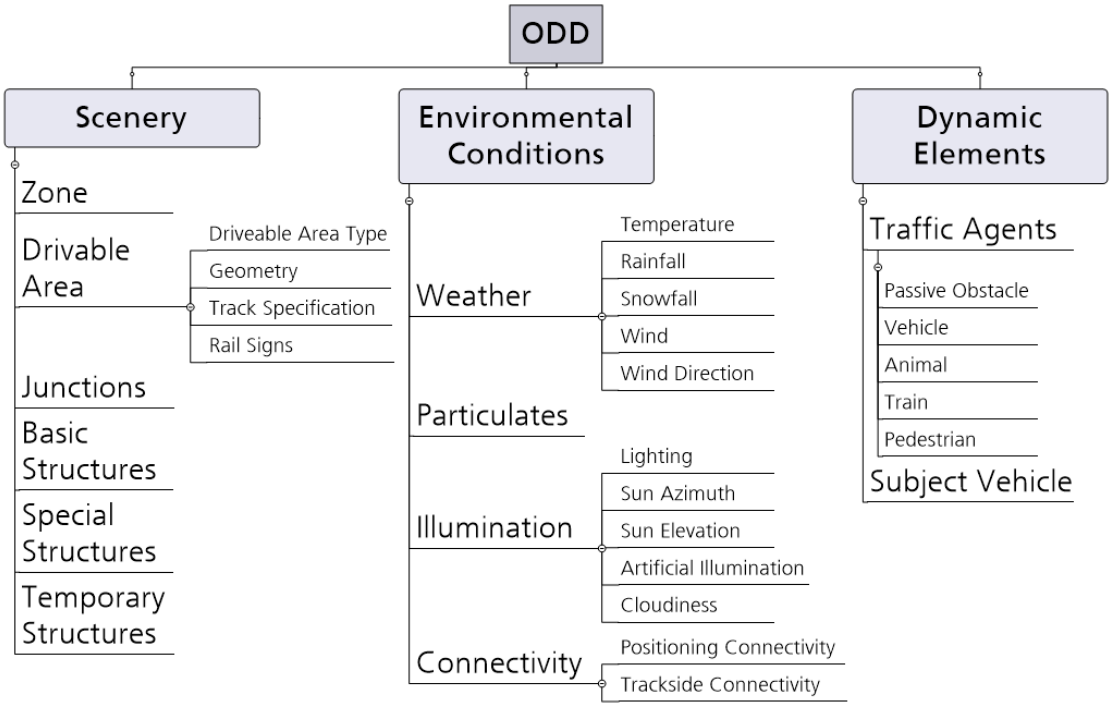


Abbildung 37: Exemplarischer Auszug aus der ODD für den Bahnsektor

Darüber hinaus unterstützt die Taxonomie die Interessengruppen, einschließlich der Regulierungsbehörden, bei der Verständigung und Überprüfung der genauen

Betriebsbedingungen und stellt so die Einhaltung strenger Sicherheitsstandards sicher. Der standardisierte Ansatz fördert eine bessere Kommunikation zwischen den verschiedenen Beteiligten, die an der Systementwicklung, Bewertung und Zertifizierungsprozessen beteiligt sind.

Die im Projekt entwickelte Taxonomie für die Operational Design Domain von vollautomatisierten Schienenfahrzeugen wurde anschließend in die Standardisierung als DIN DKE Spezifikation (DIN DKE SPEC 99004) überführt.

2.1.3.3 Architektur

Die Fahrzeuganforderungsspezifikation bildet die Grundlage für die Fahrzeugarchitektur. Diese von Siemens Mobility entwickelte Fahrzeugarchitektur umfasst innerhalb der Systemgrenzen die grundlegenden Schnittstellen zu den Teilsystemen und die darin festgelegten Funktionen. Zusätzlich zu der definierten Fahrzeugarchitektur für ein KI-basiertes Perzeptionssystem wurden Festlegungen für die Umweltwahrnehmung und der Ablauf der Fahrwegsüberwachung auf Fahrzeugebene definiert.

Entgegen dem ursprünglichen Vorhaben wurde auf die RAM-Analyse inklusive einer Instandhaltungs- und Wartungsstrategie verzichtet. Bei den Diskussionen zur Fahrzeugarchitektur hat sich herausgestellt, dass abhängig von den betrieblichen Vorgaben und den Bedingungen der Umgebung eine große Variabilität im Umfang und der Art der Sensorik besteht. Somit kann eine valide Bewertung einer solchen Strategie unter diesen Voraussetzungen nicht erbracht werden.

Ausgehend von der Architektur auf Fahrzeug-Ebene wurde von Siemens Mobility eine Architektur für ein sicheres Perzeptionssystems (System-Ebene) konzipiert und entwickelt, welches ein wesentliches Teilsystem eines voll-automatisierten Fahrbetriebes ist. Das Ziel dabei war die Spezifikation und Dokumentation der Architektur für die Objekterkennung und die Abstimmung mit anderen Fachexperten (Safety- und KI-Experten, Gutachter) innerhalb des Konsortiums.

Die Architektur beschreibt die Konzepte, Strategien und Maßnahmen zur Erreichung der funktionalen und der Sicherheitsziele, um einen fahrerlosen, voll-automatischen Fahrbetrieb (GoA3/4) auf Fahrzeugseite zu realisieren. Die Basis der Architektur des sicheren Objekterkennungssystems bilden die folgenden Anwendungsfälle (Use Cases), die aus den Anforderungen der übergeordneten Fahrzeugebene abgeleitet wurden (siehe Kapitel 2.1.3.1).

Wesentliche Aktivitäten zur Erarbeitung der System-Architektur für eine sichere Objekterkennung waren:

- die Erstellung einer High-Level Architektur,
- die Strukturierung der High-Level Architektur in Komponenten,
- die Definition von Schnittstellen,
- die Erweiterung der Architektur um die Maßnahmen, welche im Rahmen von Sicherheitsanalysen (System Level FMEA = Fehlermöglichkeits- und Einflussanalyse) erarbeitet wurden,
- die Zuweisung von funktionalen und Sicherheitszielen an die Komponenten,
- die Erweiterung der Architektur für mehrere Ausprägungsvarianten,
- die Analyse der einzelnen Architekturvarianten im Hinblick,
- die Kommunikation der Architektur und systematischer Dialog auf allen Ebenen mit den verschiedenen Fachexperten innerhalb des Konsortiums,
- die Validierung der Architektur mittels eine MVP-Ansatzes (Minimum Viable Product).

Die Architektur wurde von Siemens Mobility in dem Tool Magic Draw und mithilfe der Beschreibungssprache SysML (Systems Modeling Language) modelliert. Dabei wurden die Richtlinien aus der Siemens Mobility angewandt. Die Modellierung mit SysML ermöglichte eine integrierte und konsistente Beschreibung der Architektur ausgehend von den Anforderungen hinzu den Sicherheitsanalysen und den Komponentenbeschreibungen.

Die High-Level-Architektur (siehe Abbildung 38) stützt sich auf das seit Jahren etablierte JDL-Pattern (Joint Directors of Laboratories) [Steinberg e.a., Revising the JDL model, Proceedings of the SPIE, 1998]. Sie spezifiziert und strukturiert – auf einer abstrakten Ebene – die Teilfunktionen der Objekterkennung:

- Detect,
- Build Scene,
- Analyze Scene,
- Assess Situation,
- Monitoring.

Diese abstrakte Beschreibung definiert die erste Strukturebene und lässt gleichzeitig Raum für verschieden Lösungsvarianten, so dass die verschiedenen Forschungsansätze der einzelnen Konsortialpartner in die Realisierung des Objekterkennungssystems einfließen konnten. Die High-Level Architektur mit ihren Architekturschnitten entlang der Datenverarbeitung stellt zudem die Basis für eine Modularisierung des Systems bereit.

Die High-Level Architektur wurde weiterhin in die europäische Normierung eingebracht und dort als Referenzarchitektur akzeptiert (ERJU R2DATO, Workpackage 6.3 Architecture).

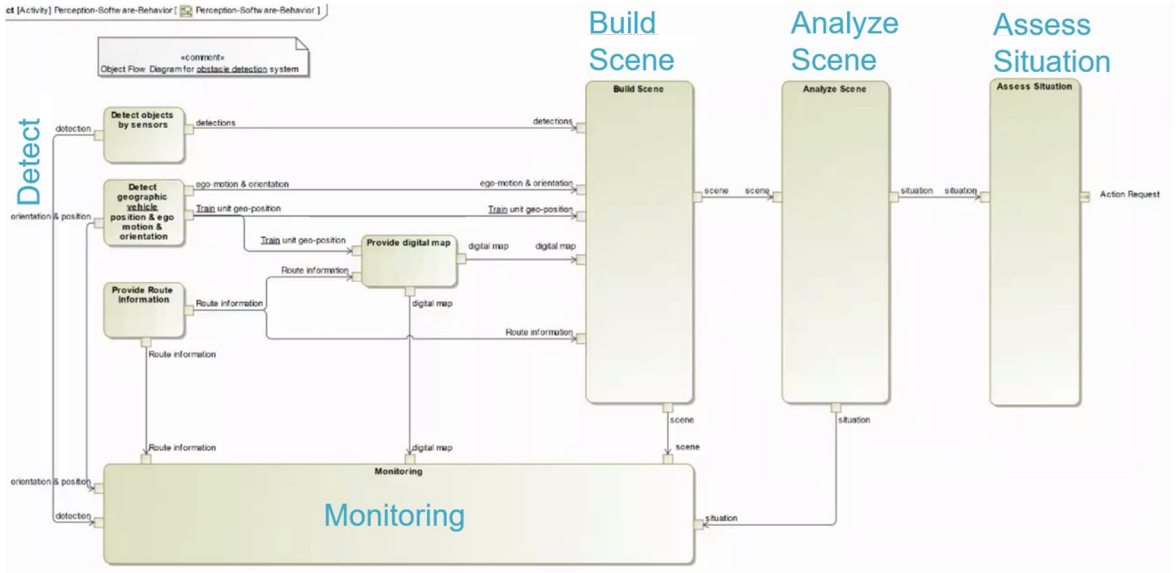


Abbildung 38: High-Level Architektur des sicheren Objekterkennungssystems

Die in der High-Level Architektur der Perzeption durch die Funktions-Blöcke „Build Scene“ und „Monitoring“ verwendete digitale Karte für die im Projekt definierte Teststrecke auf dem Gelände der Havelländischen Eisenbahn (HVLE) wurde seitens Siemens Mobility erstellt und an die Akteure der Funktions-Blöcke verteilt. Im Rahmen der Kartenerstellung wurde die Lokalisierung von Bahninfrastruktureinrichtungen und insbesondere die Erkennung der Streckentopologie erprobt. Die Herausforderungen bei der Erstellung der Digitalen Karte betrafen insbesondere das Zusammenführen („mergen“) der Gleismitteachsen in den Weichen- und Kreuzungsbereichen.

Als zweite Strukturierungsebene der System-Architektur wurde ein Komponentendiagramm als Zerlegung und Verfeinerung der High-Level Architektur erstellt. Die einzelnen Komponenten wurde den einzelnen Ebenen und Teilfunktionen der High-Level Architektur zugeordnet. Weiterhin wurden die Schnittstellen zwischen den einzelnen Komponenten definiert, wodurch eine Austauschbarkeit von Komponenten ermöglicht wird.

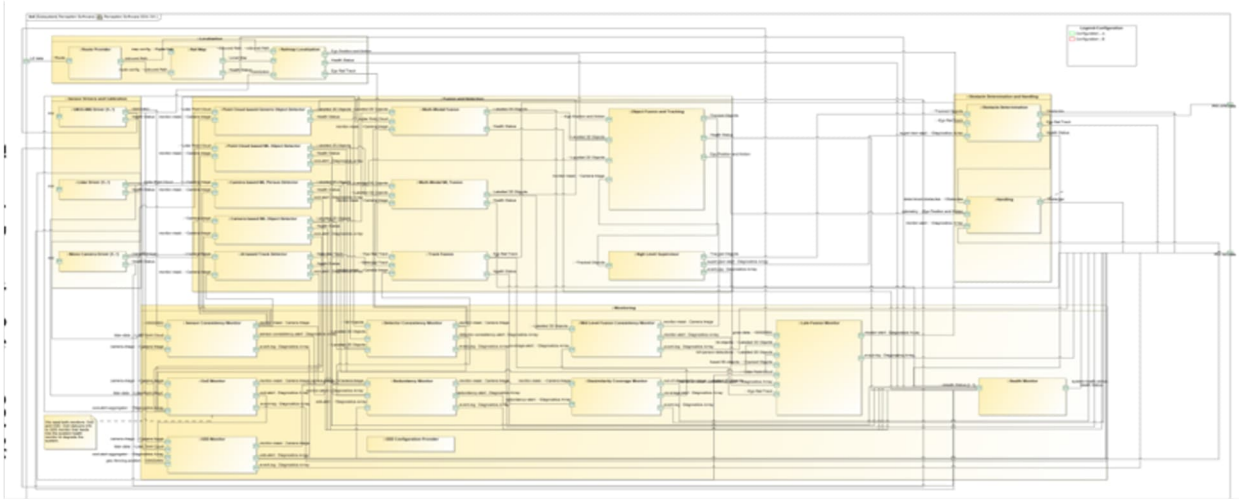


Abbildung 39: Architektur auf Komponenten-Ebene (aus Vertraulichkeits-Gründen unscharf dargestellt)

Das Komponentendiagramm ist als sogenanntes 150% Modell erstellt, welche verschiedene Systemvarianten ermöglicht. Dabei sind nicht alle Komponenten und Pfade durch das System gleichzeitig aktiv, sondern nur eine gezielte, konfigurierbare Auswahl. Dadurch wurde eine Evaluation verschiedener Algorithmen zur Objekterkennung und Sensorfusion ermöglicht.

Für den systematischen und strukturierter Dialog mit den Komponentenverantwortlichen (z.B. für den Personendetektor oder für die High-Level Fusion) wurde ein Architekturfragebogen erstellt, welche die wesentlichen Aspekte für die Ausarbeitung dieser Komponenten umfasst:

- Lösungsstrategie,
- Beiträge zur Erfüllung der funktionalen und Sicherheitsanforderungen,
- Berücksichtigung bzw. Beeinflussung der Landscape of AI Safety Concerns,
- Unterstützung der spezifizierten Sicherheits-Maßnahmen,
- Fehlermodi der Komponenten und inne liegenden Algorithmen,
- Vor- und Nachbedingungen der Komponenten für die Verarbeitung der Daten.

Mithilfe des Architekturfragebogen wurden detaillierte Antworten der Komponentenverantwortlichen gewonnen, welche die Basis waren für:

- die gezielte Zusammenstellung von einzelnen Datenpfaden,
- die Analyse dieser Datenpfade und der gesamten Architektur,
- eine verlässliche Zuordnung der verschiedenen Anforderungen und Sicherheits-Maßnahmen auf diese Komponenten,
- eine Identifikation offener Lücken in der Systemrealisierung.

Die Fragebögen bilden einen wichtigen Baustein für das (Sicherheits-) Architektur-Dokuments, welches die Architektur-Konzepte, -Strategien und -Maßnahmen für die Erreichung des Sicherheitsziels beschreibt.

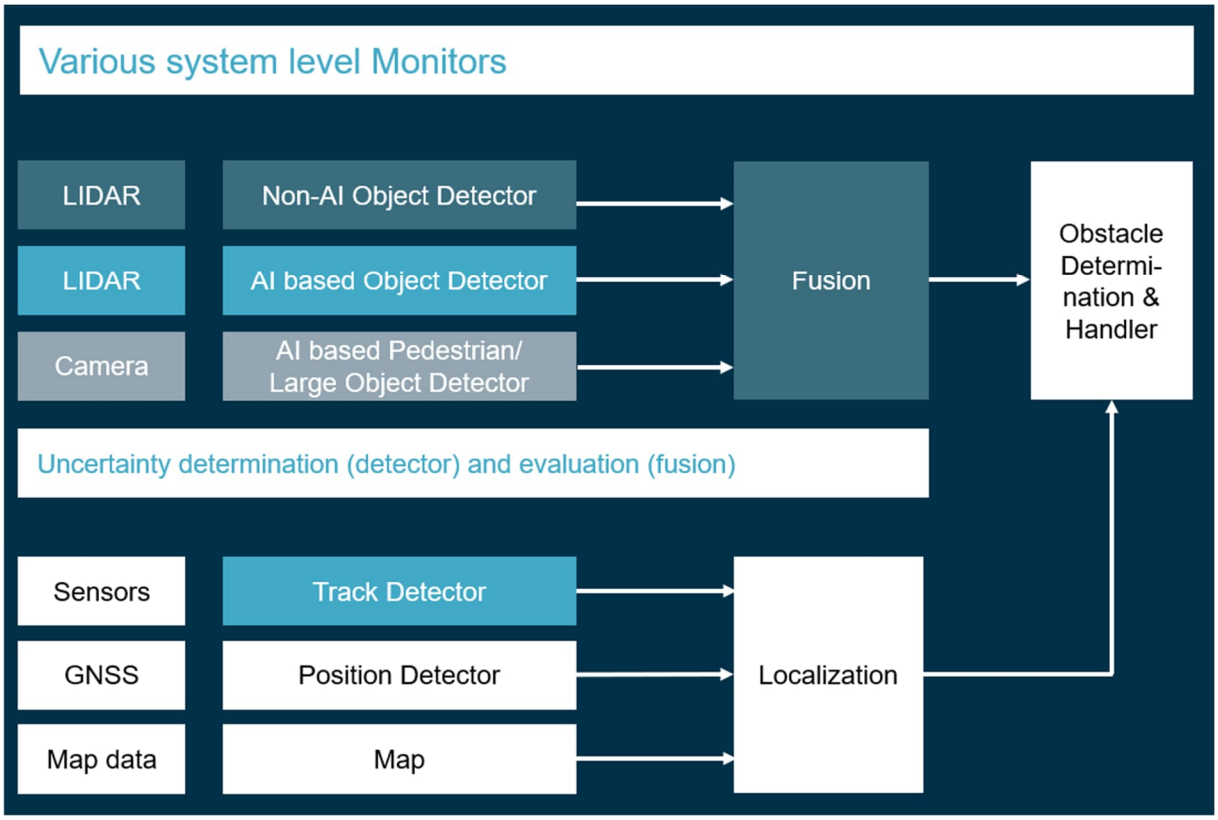


Abbildung 40: Prinzipien der Architektur für eine sichere Objekterkennung

In Zusammenarbeit mit den Safety-Experten, die sich mit der Sicherheitsnachweisführung beschäftigten, wurden Prinzipien erarbeitet und auf die Architektur angewandt, welche die Architektur der Objekterkennung zu einer sichereren Architektur weiterentwickeln (siehe Abbildung 40). Diese Architekturprinzipien umfassen:

- Definition dissimilarer Architekturelemente und Datenpfade unter Verwendung von verschiedenen Sensormodalitäten und verschiedenen Detektoren mit und ohne KI-Algorithmen:
  Die im safe.trAIn-Projekt entwickelte Sicherheitsarchitektur integriert explizit verschiedene Sensormodalitäten und Detektionsmethoden, die sowohl KI-basierte als auch konventionelle nicht-KI-Detektionspfade umfassen. Dieser duale Ansatz gewährleistet Robustheit und Redundanz, die für das zuverlässige und sichere vollautomatisierte Fahren entscheidend sind.

- Unsicherheitsbestimmung und -propagation durch diese Datenpfade ausgehend von den Sensormessungen hinzu der Sensorfusion:

Ein Kernprinzip dieser Architektur ist der systematische Umgang mit Unsicherheiten durch integrierte Mechanismen zur Bestimmung und Weitergabe von Unsicherheiten. Diese Mechanismen bewerten das Vertrauen in einzelne Sensorwerte und die daraus resultierenden Fusionsentscheidungen, wodurch die Gesamtzuverlässigkeit der Systementscheidungen verbessert wird.

- Überwachung des Systems und der Komponenten zur Laufzeit mithilfe intrinsischer und extrinsischer Monitore:
  Die safe.trAIn-Architektur umfasst ein umfassendes Systemmonitoring, das explizit darauf ausgelegt ist, die Anforderungen an Sicherheitsnachweise zu erfüllen. Dieses Monitoring überprüft kontinuierlich die Betriebsleistung, erkennt Abweichungen umgehend und stellt sicher, dass sicherheitskritische Funktionen aufrechterhalten werden. Die Integration dieser Prinzipien stellt sicher, dass das System nicht nur strenge Sicherheitsstandards erfüllt, sondern auch klare und transparente Dokumentationen zur Unterstützung der behördlichen Vorschriften und des Vertrauens der Interessengruppen bereitstellt.

Die einzelnen Komponenten des Systems wurden so zu Datenpfaden kombiniert, dass diese Datenpfade spezifische Eigenschaften und Teilaufgaben des Systems übernehmen (z.B. Performance Datenpfad, erklärbarer Datenpfad). Für die Analyse dieser Datenpfade wurde eine Auswahl von Kriterien definiert (analog zum Architekturfragebogen für die Komponenten); in Zusammenarbeit mit den Fachexperten wurden dann die Datenpfade anhand dieser Kriterien analysiert und dokumentiert. Im Ergebnis entstand eine Architektur, welche durch eine gezielte Kombination der Datenpfade eine spezifische, an die Domäne anpassbare Systemrealisierung aufweist. Die Analysierbarkeit des Systems war zudem ein wichtiger Beitrag für das Erstellen des Sicherheitsnachweises.

Im Ergebnis entstand eine Architektur für das Objekterkennungssystem, welche die funktionalen und Sicherheitsziele im Projekt safe.trAIn nachvollziehbar realisiert. Die Beiträge der einzelnen Partner (z.B. Algorithmen und Komponenten zur Detektion mit künstlicher Intelligenz oder Komponenten zur Sensorfusion) werden in ein konsistentes und integriertes Architekturmodell abgebildet. Die Architektur ist damit die Basis für die Analyse der Systemfähigkeiten, für die Entwicklung der verschiedenen Systemrealisierungen für das virtuellen Testfeld und die Sicherheitsnachweisführung. Die finalen Ergebnisse der Architekturarbeit sind in einem detaillierten Architekturdokument zusammengefasst.

Zur Entwicklung der System-Architektur gehört auch die Erarbeitung von Architekturmustern für die sichere Sensor-Fusion und deren Integration, Anwendung und Evaluation innerhalb des Perzeptionssystems.

Die grundsätzliche Aufgabe der Sensorfusion ist die Integration der Informationen aus allen Sensoren in ein gemeinsames Umgebungsmodell (Umfeldmodell), welches die Informationen über die Objekte aus der statischen und dynamischen Umgebung (Positionen, Größe, Objektklasse etc.), des eigenen Schienenfahrzeugs (Position, Geschwindigkeit) und des weiteren Schienenverlaufs enthält. Dieses Umfeldmodell stellt eine digitale Abbildung (Zwilling) der Umgebung dar; durch die zeitlich und örtlich synchronisierte Darstellung aller Informationen in einem Modell lassen sich dann konsistente Entscheidung treffen, zum Beispiel dass ein Umgebungsobjekt ein Hindernis darstellt und dass aufgrund der aktuellen Ego-Fahrgeschwindigkeit eine Bremsauslösung erfolgen muss.

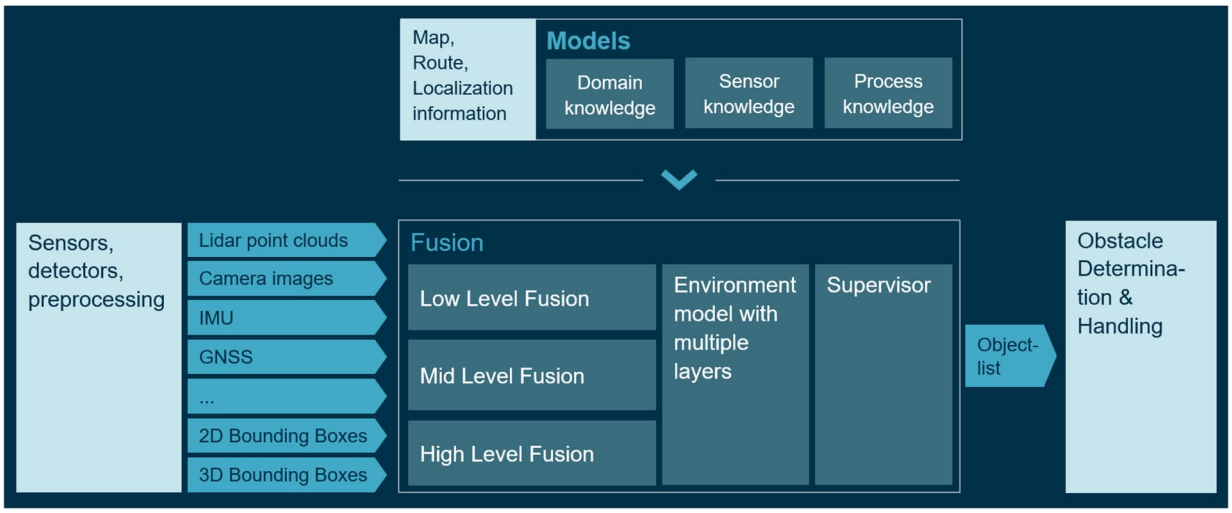


Abbildung 41: Konzepte in der Architektur der sicheren Sensorfusion

Als Basis für die Architekturarbeit der Sensorfusion wurden die folgenden Prinzipien der übergeordneten Architektur des Objekterkennungssystems angewandt und durch die Sensorfusion realisiert:

- Detektion von Objekten mit dissimilaren Pfaden (Sensoren, Detektoren),
- Unsicherheitspropagation durch das System,
- Monitoring.

Die Abbildung 41 stellt die verschiedenen Konzepte der Architektur der Sensorfusion dar, deren Zusammenwirken innerhalb des Konsortiums zusammen mit der Siemens AG erarbeitet wurden.

Für die Fusion der einzelnen Eingangsinformationen der Sensoren und Detektoren wurde ein Konzept einer kaskadierten Fusion entwickelt:

- Low Level Fusion: zur Fusion gleichartiger Sensormessungen, wie zum Beispiel die Punktwolken zweier Lidarsensoren,
- Mid Level Fusion: zur Fusion von Sensormessungen von zwei-dimensionaler und drei-dimensionaler Messungen, wie zum Beispiel Punktwolken aus Lidarsensoren und Bounding Boxen aus Kameradetektionen,
- High Level Fusion: zur generellen Fusion heterogener Sensormessungen auf Merkmalsebene (Detektionen) und zur Verfolgung dieser über der Zeit (Tracking).

Innerhalb der Arbeiten zur Sensorfusions-Architektur wurden durch die Siemens Mobility die Architektur für die High Level Fusion und die Konzepte für deren Monitoring sowie für den Supervisor ausgearbeitet. Im Kern der High Level Fusion wird über einen Kalman Filter die Objekthypothesen des Umfeldmodells iterativ durch neue Messungen der Sensoren und Detektoren aktualisiert (siehe Abbildung 42). Durch die fortlaufende Aktualisierung wird über der Zeit sowohl die Genauigkeit der einzelnen Attribute (wie zum Beispiel Position oder Größe eines Objekts) als auch deren Konfidenz bezüglich Existenz oder Objektklassifizierung (zum Beispiel Person oder Fahrzeug) signifikant erhöht.

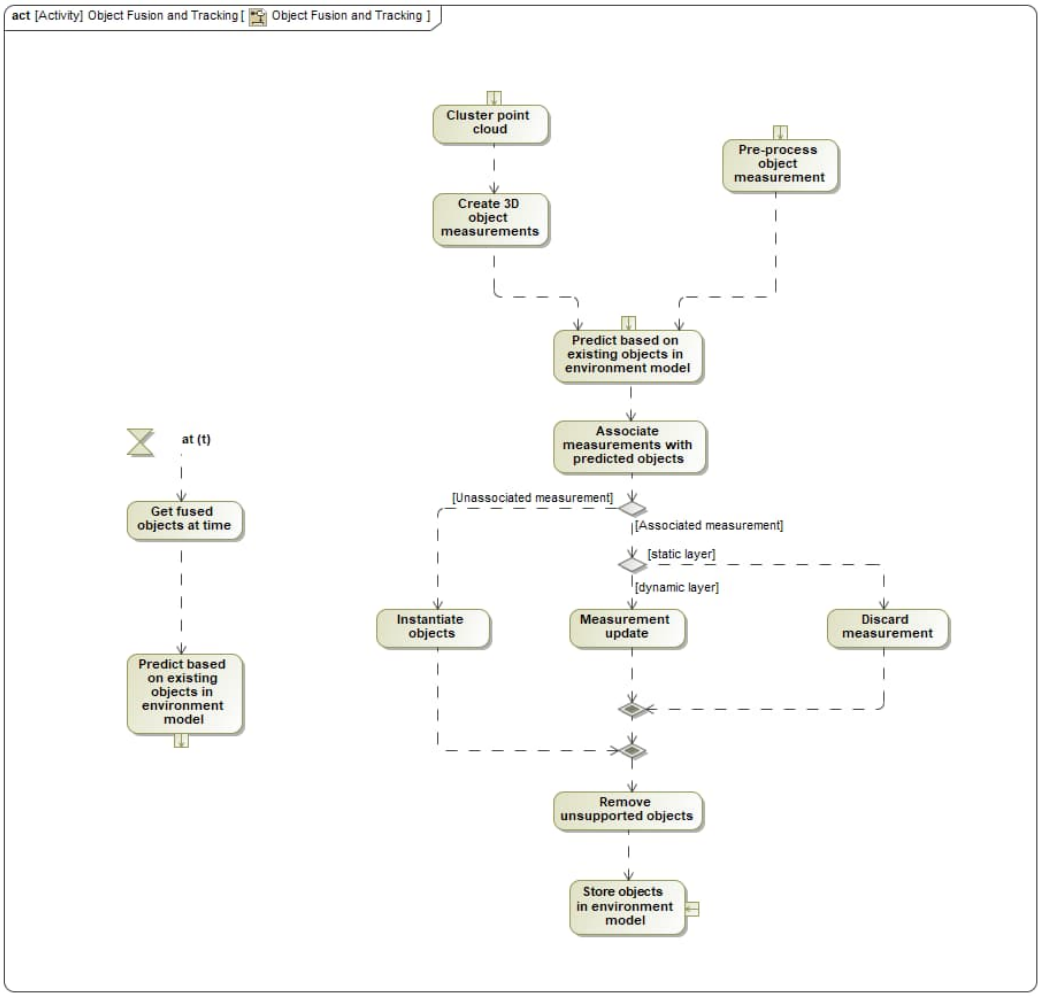


Abbildung 42: Aktivitätsdiagramm für die High Level Fusion zur Abbildung der statischen und dynamischen Umgebung in ein Umfeldmodell

Zur strukturierten Bewertung und zum Vergleich der im Rahmen des Projekts entwickelten verschiedenen Sensorpfade und Detektionsalgorithmen wurde eine sogenannte „Capability Matrix“ aufgestellt. Durch die eindeutige Dokumentation der Fähigkeiten verschiedener Sensorkombinationen unterstützt die Matrix die gezielte Auswahl von Sensoren und Algorithmen, die explizit auf die definierten Sicherheits- und Funktionsziele zugeschnitten sind. Jeder Pfad innerhalb der Matrix wird systematisch über mehrere Dimensionen hinweg bewertet.

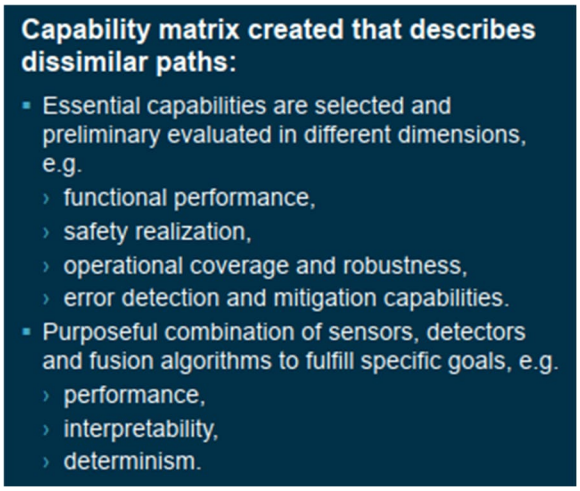


Abbildung 43: Capability Matrix zur strukturierten Beschreibung der dissimilaren Pfade

Ein entscheidender Aspekt, der in der Capability Matrix hervorgehoben wird, ist die Erklärbarkeit jeder Komponente und jedes Pfades. Klare Erklärbarkeit und Transparenz von Systementscheidungen sind unerlässlich, um belastbare und vertretbare Sicherheitsargumente zu schaffen.

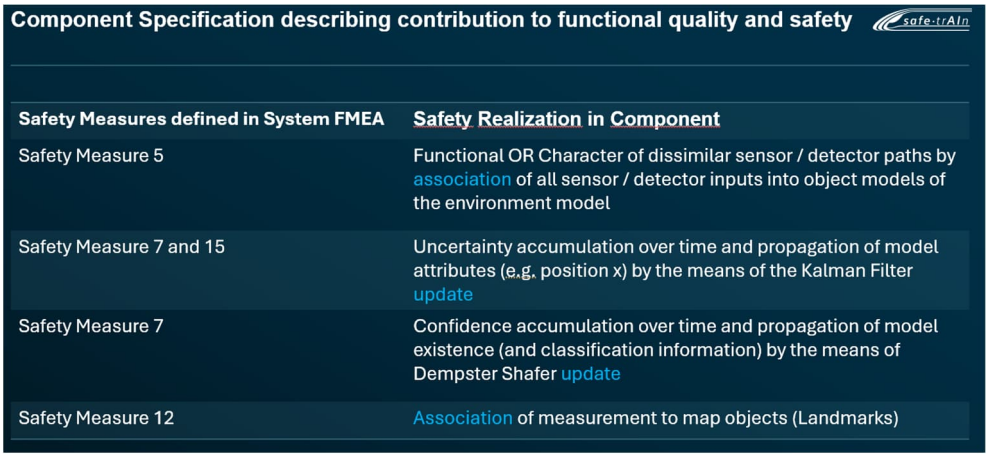


Abbildung 44: Realisierte Sicherheits-Maßnahmen durch die High Level Fusion

Mithilfe der High Level Fusion werden verschiedene Sicherheits-Maßnahmen erfüllt, welche an das übergeordnete Perzeptionsssystem gestellt werden (siehe Abbildung 44). Dazu zählen auch die Fusion der Unsicherheiten der einzelnen Detektorpfade sowohl auf der Merkmalsebene (wie zum Beispiel Position) mithilfe des Kalman Filters als auch auf Objektebene (Existenz und Objektklasse) mithilfe des Dempster Shafer Algorithmus.

Durch eine Kombination von UND-Logik (mehrere Sensoren müssen ein Objekt gesehen haben) und ODER-Logik (mindestens ein Sensor muss ein Objekt gesehen haben) ermöglicht

die erarbeitete High Level Fusion eine Balance zwischen Sicherheit (hohe Detektionsraten) und Verfügbarkeit (niedrige Falschalarmraten).

Zur Absicherung der High Level Fusion selbst, wurden zwei Konzepte entwickelt:

- Late Fusion Monitor, welcher als extrinsischer Monitor die Nachbedingungen des High Level Fusion überwacht,
- High Level Supervisor, welcher als intrinsischer Monitor im Datenpfad die ausgegebene Objektliste auf semantische und Kontinuitätsbedingungen überwacht.

Mit diesen beiden Monitoren wird die High Level Fusion als Teil des Objekterkennungssystems zusätzlich abgesichert.

Im Ergebnis entstanden Architekturmuster und Konzepte für eine kaskadierte Sensorfusion, welche in die Architektur des Objekterkennungssystems integriert wurden und welche einen zentralen Baustein für die Erreichung einer sicheren Objekterkennung darstellen. Die Architekturdokumentation der Sensorfusion und ihrer Komponenten sind in die Architekturdokumentation des Objekterkennungssystems eingeflossen. Die Ergebnisse der Sensor-Fusions-Architektur sind wesentlicher Input für die Entwicklung des sicheren Objekterkennungssystems, dessen Bewertung über Metriken im virtuellen Testfeld und für die Erstellung von Sicherheitsanalysen und -nachweisen.

2.1.3.4 Implementierung

Basierend auf den Vorarbeiten aus AP3, insbesondere der entwickelten Architektur, wurden die Komponenten eines KI-basierten Perzeptionssystem implementiert und integriert. Dies bildete das prototypische Testsystem bzw. das System under Test (SuT), welches anschließend im Virtuellen Testfeld mithilfe von Metriken bewertet wurde.

Das im Projekt entwickelte SuT umfasst folgende Komponenten:

- KI-basierte Gleis-Detektion (Track Detektor),
- Verschiedene KI-basierte Ansätze zur Personen-Detektion,
- KI-basierte Detektion von großen Objekten (Large Obstacle Detektor),
- Verschiedene Ansätze für die Fusion von Sensordaten.

Die von Siemens Mobility entwickelte kamera-basierte Gleis-Detektion nutzt Kamerabilder, um eine genaue Lokalisierung und detaillierte Kartierung von Bahngleisen/-strecken zu

generieren. Durch die Einbeziehung von Kamera-Kalibrierungsinformationen und bahnspezifischen Einschränkungen kann das System räumliche Informationen, die für eine sichere Zugnavigation unerlässlich sind, genau interpretieren und modellieren. Dieser Ansatz kombiniert fortschrittliche 2D-Bildsegmentierungstechniken mit 3D-Mapping, um eine umfassende und präzise Darstellung der Strecke und ihrer Umgebung zu erstellen. Eine solche detaillierte Gleis-Detektion ist entscheidend, um das Lichtraumprofil („Clearance Space“), den Warn- und den Gefahrenbereich (siehe Kapitel 2.1.3.1) vor dem Fahrzeug zu ermitteln, um daraus abzuleiten, ob sich Objekte im Fahrschlauch des Zuges befinden und ein potenzielles Hindernis darstellen.

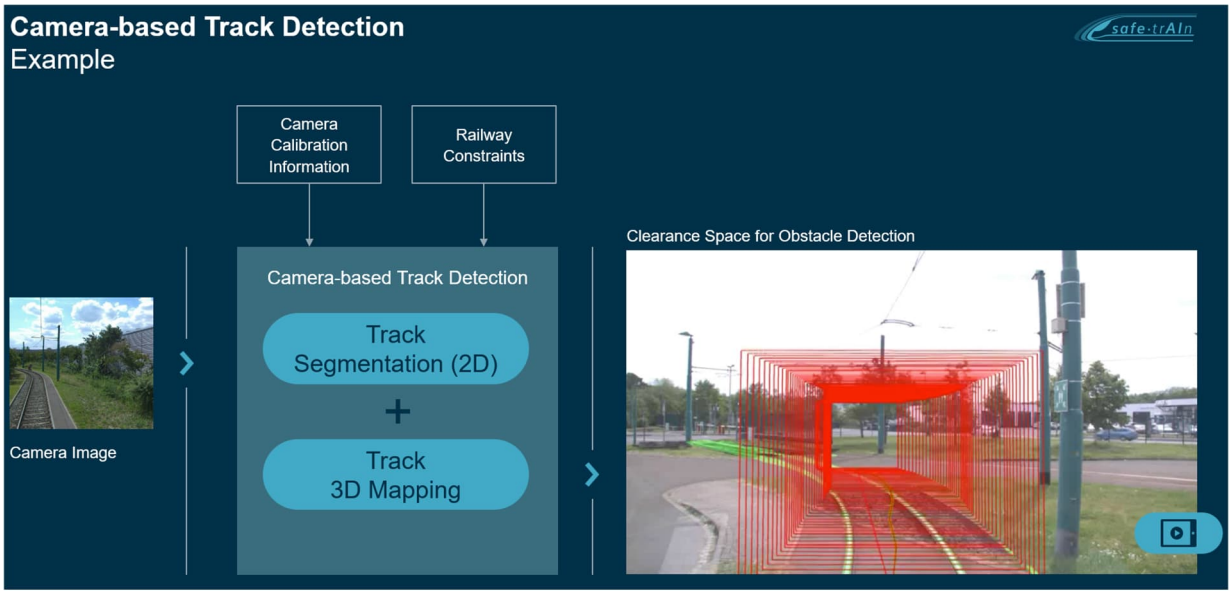


Abbildung 45: Kamera-basierte Gleiserkennung

Ein Beispiel für die implementierte Personenerkennung ist der Panoptic Segmentation Detector, den Siemens Mobility ins Projekt eingebracht hat. Die panoptische Segmentierung ist eine fortschrittliche Deep-Learning-Technik, die ein detailliertes semantisches Verständnis des Bildinhalts ermöglicht. Die panoptische Segmentierung klassifiziert gleichzeitig jedes Pixel in einem Bild und identifiziert gleichzeitig einzelne Objekte eindeutig. Sie eignet sich daher besonders für die präzise und robuste Fußgängererkennung in komplexen Bahnumgebungen. Aktive Lernmethoden werden speziell in Datensätze aus dem Schienenfahrzeug-Umfeld integriert, so dass das Modell schrittweise aus den Szenarien lernen und seine Erkennungsfähigkeiten kontinuierlich verbessern kann. Durch die Feinabstimmung wird das Segmentierungsmodell hochgradig optimiert, um Fußgänger auch in unterschiedlichen und komplexen Kontexten genau und zuverlässig zu erkennen.

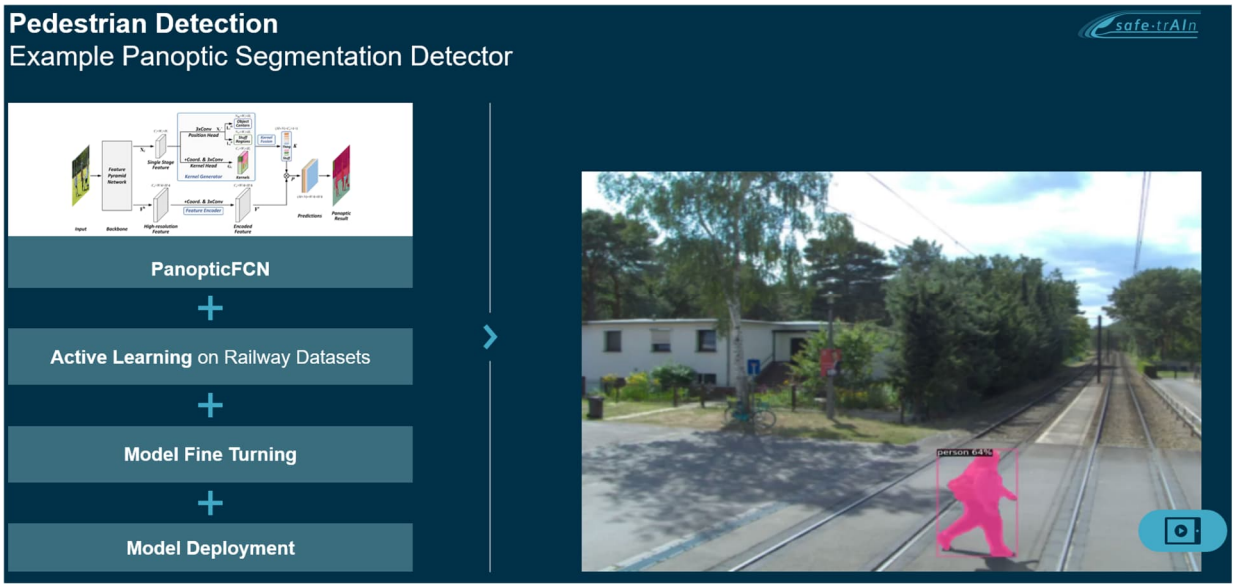


Abbildung 46: Personen-Detektion basieren auf PanopticFCN Netzwerk

Darüber hinaus hat Siemens Mobility einen weiteren KI-Algorithmus für die Detektion von Personen entwickelt. Dieser basiert auf RBF-Netzwerken (Radialen Basis Funktionen), welcher aufgrund der Beschaffenheit des Netzes (nur 3 Netzwerk-Layer: Eingabe-Schicht, versteckte Schicht, Ausgabe-Schicht) eine gute Erklärbarkeit vorweist. Ein RBF-Netzwerk funktioniert, indem es Eingabedaten mithilfe radialer Basisfunktionen in einen höherdimensionalen Raum transformiert, wo es durch gewichtete Summen dieser Funktionen eine nichtlineare Approximation oder Klassifikation ermöglicht.

Um den zweiten betrachteten Use Case (Große Hindernisse) abzudecken, wurde seitens Siemens Mobility im Projekt ein auf Foundation Modellen basierender „Large Obstacle Detektor“ entwickelt.

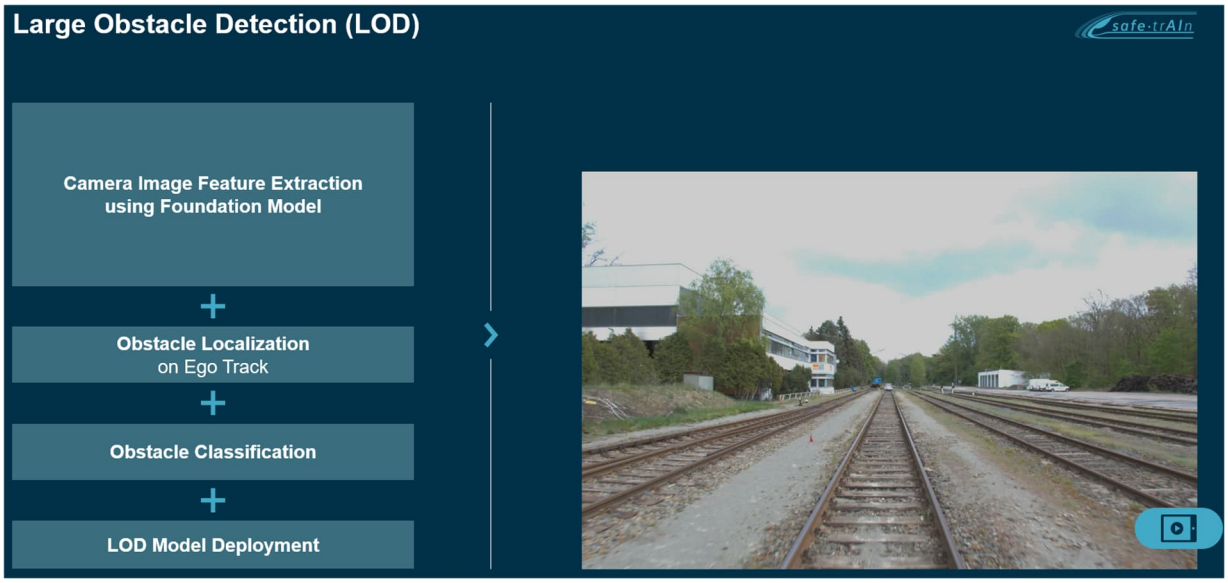


Abbildung 47: Detektor für große Hindernisse

Ein wesentlicher Bestandteil des Systems ist die Fusion. Die in der Architektur angedachten High-Level Fusions-Ansätze (siehe Kapitel 2.1.3.3) wurden im System under Test seitens Siemens Mobility prototypisch implementiert, während die die Konzeption und Implementierung der Mid Level Fusion durch den Projektpartner Siemens AG erfolgte.

Für das Training der verschiedenen KI-Algorithmen kamen sowohl frei zugängliche Datensätze wie OSDaR23 (Open Sensor Data for Rail 2023) als auch von Siemens Mobility erzeugte Daten von vorherigen Forschungsaktivitäten (wie z.B. Advanced TrainLab) zum Einsatz, welche dem Konsortium zu Trainingszwecken zu Verfügung gestellt wurden.

2.1.3.5 Sicherheitsnachweis für die GOA3/4 Architektur

Basierend auf der Architektur für ein KI-basiertes Perzeptionssystem wurde die Struktur und Argumentationslinie des Sicherheitsnachweises erarbeitet und die für die Validierung des Systems notwendigen Evidenzen und Nachweise an das Training und Testen festgelegt.

Die im Projekt, maßgeblich durch Siemens Mobility und Siemens AG, entwickelte Sicherheitsnachweis-Strategie basiert auf 5 Säulen:

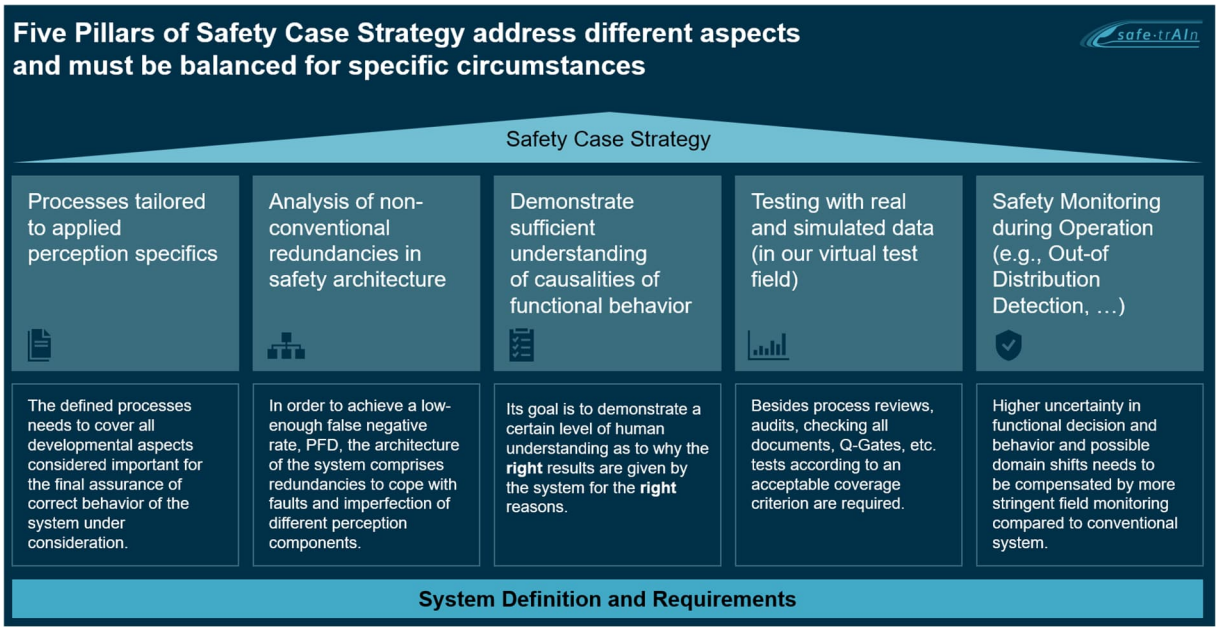


Abbildung 48: 5 Säulen der Sicherheitsnachweis-Strategie

- Prozesse:
  Der Entwicklungsprozess muss speziell auf die Besonderheiten der angewandten Perzeptions-Aufgabe abgestimmt sein und alle für die finale Sicherheitsargumentation nötigen Aspekte berücksichtigen. Hierbei spielt die ODD (siehe Kapitel 2.1.3.2) als zentrales Element im Entwicklungsprozess ebenso wie die Landscape of AI Safety Concerns (siehe Kapitel 2.1.2) eine entscheidende Rolle. Um KI-Systeme sicher zu gestalten, müssen alle Sicherheitsbedenken (AI Safety Concerns) adressiert werden. Für jedes einzelne Bedenken sind über den gesamten Entwicklungsprozess und Lebenszyklus der KI hinweg überzeugende Nachweise erforderlich, die belegen, dass eine ausreichende Risikominderung erfolgt ist.

- Analyse von nicht-konventionellen Redundanzen:
  Die Sicherheitsarchitektur (siehe auch Kapitel 2.1.3.). für ein KI-basiertes Perzeptionssystem soll innovative, nicht-konventionelle Redundanzen integrieren, um die Robustheit und Sicherheit des Systems zu verbessern. Im Gegensatz zu herkömmlichen Sicherheitsansätzen, die ausschließlich auf identische redundante Komponenten setzen, verwendet diese Strategie mehrere Sensormodalitäten (z. B. LIDAR, Kameras), verschiedene Erkennungsmethoden (KI-basierte und konventionelle Algorithmen) und unterschiedliche Daten-Verarbeitungs-Pfade, um eine umfassende Abdeckung und Fehlertoleranz gegenüber unterschiedlichen Ausfällen zu gewährleisten (Konzept der Dissimilarität).

Ein kritischer Aspekt dieser Redundanzstrategie ist die Bestimmung und Bewertung von Unsicherheiten, um sicherzustellen, dass die Vertrauensniveaus jedes Sensors und jedes Erkennungspfades kontinuierlich bewertet und verwaltet werden. Dieser umfassende Überwachungsansatz (Konzept der Montitore) schützt sowohl KI-gesteuerte als auch konventionelle Erkennungskomponenten und stellt sicher, dass sie effektiv zusammenarbeiten.

- Verständnis der Kausalitäten des funktionalen Verhaltens:
  Hierbei ist sicherzustellen, dass das System die richtigen Dinge aus den richtigen Gründen tut. Die enge Zusammenarbeit von KI- und Domänenexperten ist hierbei entscheidend. Durch das Schaffen von Transparenz und Vertrauen in die Entscheidungsfindung des Systems kann ein ausreichendes Verständnis der Kausalzusammenhänge hinter dem funktionalen Verhalten des Systems nachgewiesen werden. Hierbei liegt der Fokus darauf, zu analysieren, warum das System bestimmte Entscheidungen trifft – und nicht nur, welche Entscheidungen es trifft. Dazu gehört, soweit möglich, auch die Identifikation potenzieller Voreingenommenheiten (Biases) oder verzerrender Einflussfaktoren (Confounder).
  Eine vollständige End-to-End-Erklärbarkeit ist nicht realistisch. Dieses Prinzip fordert daher, auf Komponentenebene ein angemessenes Maß an Beobachtbarkeit und Erklärbarkeit bereitzustellen. Dies kann mithilfe von Techniken wie z. B. Saliency Maps (siehe Kapitel 2.1.2.) erreicht werden.
  Das Verständnis der Kausalitäten des funktionalen Verhaltens verbessert nicht nur das Vertrauen und die Akzeptanz der Stakeholder, sondern hilft auch dabei, potenzielle Verzerrungen oder Störfaktoren zu identifizieren und zu mindern, die die Systemleistung negativ beeinflussen könnten. Letztendlich ist es für eine überzeugende Sicherheitsargumentation und den erfolgreichen Einsatz vollautomatisiert fahrender Regionalzüge unerlässlich, klare kausale Begründungen für KI-Entscheidungen nachzuweisen.

- Testen:
  Es ist nötig, ein umfassendes Testkonzept mit ausreichender Testabdeckung zu implementieren, um die für die Sicherheit des Systems nötigen Evidenzen zu generieren. Dabei konzentriert sich jede Teststufe auf ein spezifisches Testobjekt und ein entsprechendes Testziel und wird durch eine entsprechende Testumgebung unterstützt (siehe Kapitel 2.1.4.). Im Projekt wurde dies im Rahmen des virtuellen

Testfeldes durchgeführt, welches die Ausführung sowohl von Tests auf Komponenten-Ebene als auch auf Gesamtsystem-Ebene mit verschiedenen Konfigurationen erlaubt. Es wurde sowohl mit realen als auch mit synthetischen Daten getestet.

- Sicherheitsüberwachung während des Betriebs:
  Dies umfasst die kontinuierliche Überwachung des Systems, um sicherzustellen, dass es auch während des Betriebes sicher bleibt. Die Systemüberwachung auf Ebene des Gesamtsystems ist integraler Bestandteil des in safe.trAIn entwickelten Sicherheitsnachweis-Ansatzes und bietet eine kontinuierliche, Echtzeit-Überwachung und Validierung der Systemleistung. Dieser Ansatz ist nicht nur für die anfängliche Zertifizierung, sondern auch für die Aufrechterhaltung der laufenden Betriebssicherheit unerlässlich.
  Hierzu kommen Runtime-Monitoring Methoden, wie Out-Of-Distribution (OoD) Detektion, zum Einsatz. Die Herausforderung hierbei besteht darin, zwischen gültigen OoD-Objekten und Hintergrund-Objekten zu unterscheiden, da die Auftretenswahrscheinlichkeiten von Objekten stark variieren können.

Die Sicherheitsnachweis-Strategie betont ein detailliertes Verständnis des funktionalen Verhaltens und der kausalen Mechanismen des KI-basierten Perzeptionssystems, wodurch sichergestellt wird, dass die Entscheidungen des KI-basierten Systems transparent und gerechtfertigt sind. Eine derart umfassende Sicherheitsnachweis-Strategie unterstützt die Zuverlässigkeit, Sicherheit und öffentliche Akzeptanz des vollautomatisierten Fahrens.

Ein wichtiger Beitrag zur Sicherheitsnachweis-Führung ist die Durchführung von Fehlermöglichkeits- und Einflussanalyse (FMEA) auf Komponenten-Ebene in frühen Entwicklungsphasen, um die Herausforderungen und potenziellen Ursachen von Fehlern zu verstehen, die Konsequenzen zu evaluieren und nötige Mitigations-Maßnahmen zu implementieren. Diese Analysen sind entscheidend für das Verständnis und die Integration von Sensoren, Algorithmen und Fusionsmethoden in das Wahrnehmungssystem. Im Projekt wurden maßgeblich durch Siemens Mobility und Siemens AG exemplarisch 4 FMEAs für die folgenden Komponenten durchgeführt:

- Kamera Sensor,
- Mid Level Fusion,
- High Level Fusion,
- Gleis-Detektion.

Neben den FMEAs auf Komponenten-Ebene ist auch die systematische Sicherheits-Analyse der Architektur (System Level FMEA) ein wesentlicher Baustein für die Sicherheits-

Argumentation. Mithilfe dieser konnte die Architektur um die identifizierten Sicherheits-Maßnahmen (Safety Measures) iterativ ergänzt und verfeinert werden (siehe dazu auch Kapitel 2.1.3.3).

Die wesentlichen Prinzipien, auf denen die Sicherheit-Architektur aufgebaut ist, sind

- Nicht-konventionelle Redundanzen,
- Konzept der Monitore (auf Detektor- und System-Ebene),
- Bestimmung und Bewertung von Unsicherheiten.

Dank ihrer systematischen Natur von FMEAs unterstützen diese proaktive das Risikomanagement, reduzieren potenzielle Sicherheitsrisiken und unterstützen damit die Sicherheitsargumentation für vollautomatisches Fahren von Regionalzügen.

Auf Basis der Sicherheitsnachweisstrategie konnte im Projekt auch ein formales Vorgehen für die Nachweisführung entwickelt werden. Dabei wurde zwischen System-Ebene und Komponenten-Ebene unterschieden.

1) Sicherheitsnachweis-Konzept auf System-Ebene:

   In Anlehnung an die im Bahnsektor etablierten EN 50129 wurde im Projekt das Sicherheitsnachweis-Konzept auf System-Ebene aufgebaut. Dazu war eine Ergänzung der vorgegebenen formalen Struktur um KI- / ML-Aspekte nötig.

   Folgende Punkte wurde ergänzt:

   - KI- /ML- spezifische Aspekte,
   - Ergebnisse der Komponenten-Sicherheitsnachweise,
   - Neue Annex Tabelle zu den nötigen Schritten je KI-Lebenszyklusphase.

   Insgesamt muss mit einem höheren Maß an Unsicherheit umgegangen werden. Darüber hinaus wird mit relativen Evidenzen anstelle absoluter Evidenzen gearbeitet.

   Untermauert wurde die Sicherheitsargumentation durch Erkenntnisse und Ergebnisse aus untergeordneten Dokumenten, welche alle im Projekt erstellt wurden:

   - (Sicherheits-) Architektur-Beschreibung,
   - KI-Validierungs-Report,
   - Landscape of AI Safety Concern,
   - Sicherheits-Validierungs-Konzept,
   - Validierungs-Konzept,
   - Fact Sheets der Metriken (siehe Kapitel 2.1.2).

2) Sicherheitsnachweis-Konzept auf Komponenten-Ebene:

Das Sicherheitsnachweis-Konzept auf Komponenten-Ebene wurde bestimmt durch die Gefährdungen, welche in der Landscape of AI Safety Concerns herausgearbeitet wurden. Hier kamen GSN-Bäume (Goal Structuring Notation) zur grafischen Darstellung des Sicherheitsnachweise für die KI-Komponenten zum Einsatz.

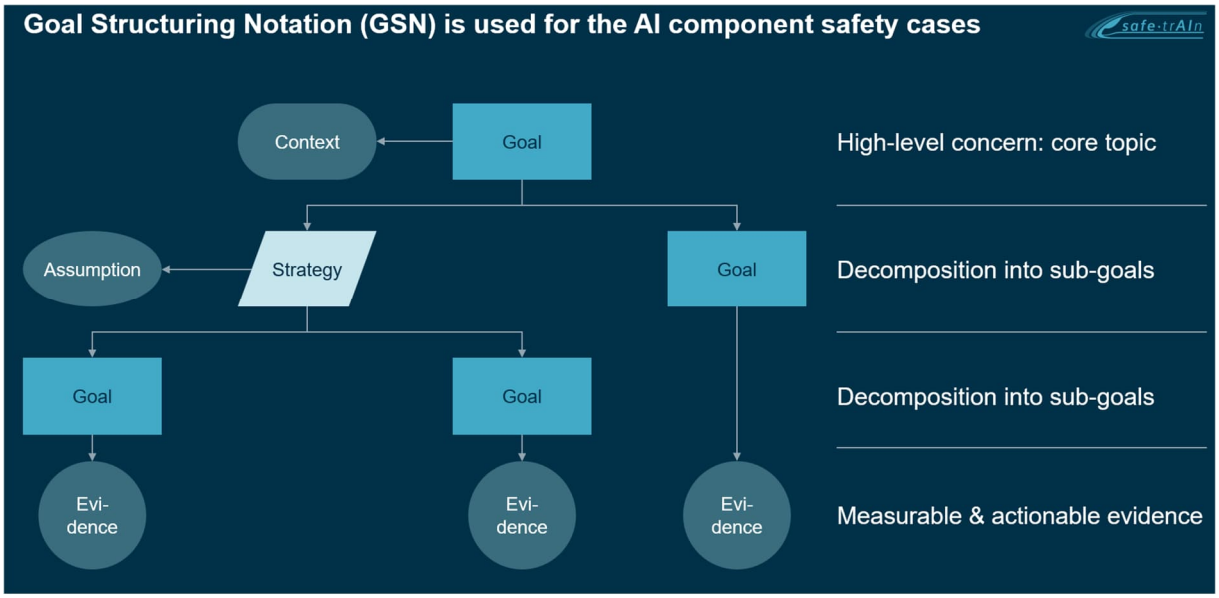


Abbildung 49: Grafische Darstellung der Sicherheitsargumentation mittels GSN

Bei der automatisierten Erzeugung von Evidenzen auf Komponenten-Ebene wurde seitens dem Partner Edge Case Research das Tool „nLoop“ im Projekt zur Verfügung gestellt und an die Siemens Mobility eigene Datenmanagementlösung (ai.store) und die von Siemens AG zur Verfügung gestellte Coding Plattform (code.siemens.com) angebunden. Damit war es möglich, Validierungs-Reporte automatisch zu generieren sowie Schwellwerte für die die Metriken anzugeben. Dies wurde im Projekt exemplarisch für die beiden KI-Detektoren (Gleis-Detektion und Mid Level Fusion) durchgeführt.

2.1.4 AP 4: Virtuelles Testfeld, Sicherheitsbewertung

In diesem Arbeitspaket wurde ein virtuelles Testfeld zur Evaluierung des entwickelten KI-basierten Perzeptionssystems und zur Generierung der für den Sicherheitsnachweis nötigen Evidenzen aufgebaut. Da reale Testfahrten im laufenden Bahnbetrieb nur eingeschränkt und mit hohem Aufwand möglich sind, setzt das Projekt auf den Einsatz eines virtuellen Testfeldes. Die KI-Komponenten können damit schon vor dem ersten physischen Einsatz umfassend validieren und verifizieren werden. Die Prüfergebnisse aus dem virtuellen Testfeld wurden mittels definierter Kriterien (Metriken aus AP2) interpretiert, in die Sicherheitsnachweis-Konzept eingebettet und durch eine Begutachtung überprüft.

Dazu gehörte im Projekt nicht nur die Konzeption und Implementierung des virtuellen Testfeldes als solches, sondern auch die Entwicklung eines kontinuierlichen Entwicklungs- und Sicherheitsprozesses (safeMLOps), die Definition der Teststrategie und Testszenarien, die Generierung von realen und synthetischen Testdaten sowie die abschließende Evaluierung und Einbettung in den Sicherheitsnachweis.

2.1.4.1 safeMLOps Prozess

Im Projekt wurde das Konzept des Continuous Integration & Contiuous Deployment („CI/CD“) verfolgt und an die Eigenheiten von KI-Anwendungen angepasst (MLOps = Machine Learning Operations). Somit verbindet dieser Workflow verbindet den traditionellen Entwicklungszyklus für Systeme (Systemengineering) mit dem spezialisierten Entwicklungszyklus für maschinelles Lernen. Der so entstandene safeMLOps Prozess ist ein kontinuierlicher Entwicklungs- und Sicherheitsprozess für sichere KI-Funktionen, bei dem jede Änderung am KI-Modell oder an den Daten sofort durch den gesamten Absicherungsprozess geführt werden muss (siehe Abbildung 50). Dieser verzahnt alle Schritte von der ODD- und Anforderungsdefinition über das Training, die virtuellen Tests bis hin zur Inbetriebnahme und dem Monitoring miteinander. Ändert sich das Modell oder ein Parameter, aktualisiert die Pipeline automatisch den betroffenen Sicherheitsnachweis. Es konnte aufgezeigt werden, dass bei einer Modellanpassung (etwa zur Verbesserung der Objekterkennung) alle zugehörigen Safety-Analysen und Testergebnisse konsistent mit aktualisiert werden können.

Die automatisierte Pipeline des safeMLOps Prozesses unterstützt kontinuierliche und schnelle Validierungszyklen, was rasche Anpassungen und iterative Verbesserungen der KI-Modelle ermöglicht.

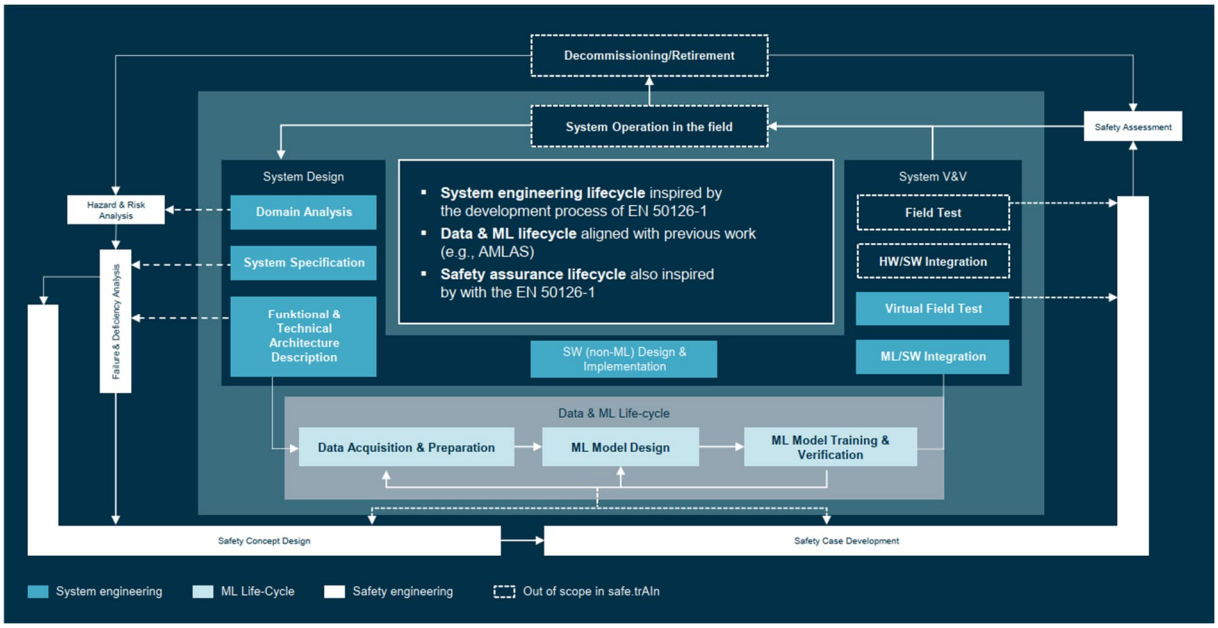


Abbildung 50: safeMLOps Prozess

Die Implementierung des safeMLOps Prozesses erfolgte mittels GitLab Runner, der von Siemens Mobility zur Verfügung gestellten Datenmanagementlösung ai.store und dem von Edge Case Research entwickelten Sicherheitsnachweis-Framework nLoop.

2.1.4.2 Teststrategie

Entgegen der ursprünglichen Vorhabens-Planung hat sich im Projektverlauf der Bedarf für eine Teststrategie herausgestellt, um das konkrete Testvorgehen eindeutig definieren zu können. Dieses Dokument umfasst sowohl die Testziele als auch die nötigen Teststufen, Testmethoden und Testmanagement-Werkzeuge.

Die Teststrategie sieht vor, dass die Tests auf verschiedenen Testebenen ablaufen.

- Komponenten-Ebene
- Integrations-Ebene
- System-Ebene

Dabei fokussiert sich jede Testebene auf ein spezifisches Testobjekt und Testziel und ist dabei mit der entsprechenden Testumgebung verknüpft (siehe Abbildung 51).

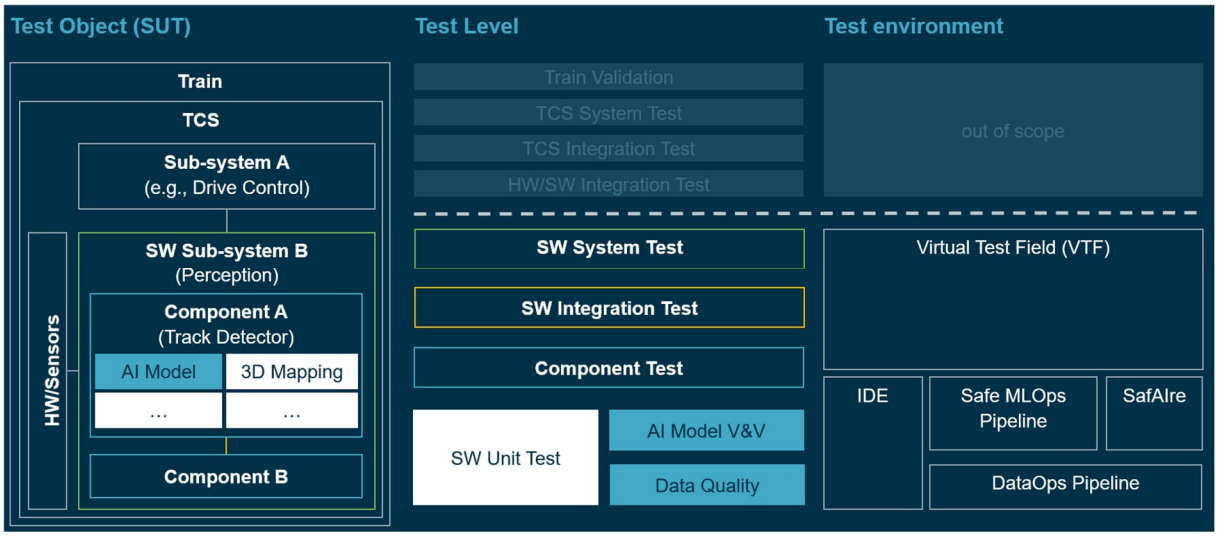


Abbildung 51: Teststrategie mit verschiedenen Testebenen

2.1.4.3 Konzeption und prototypische Umsetzung des virtuellen Testfelds

Abgeleitet von den Teststufen (Software System-, Integrations- und Komponententests) wurde im Projekt eine Test-Architektur aufgesetzt, deren automatische Durchführung der Tests mit Apache Airflow als einem Virtual Testfield Manager erfolgte. Die Test-Architektur sieht es vor, dass die Testdaten aus dem ai.store geladen und entsprechend vorverarbeitet werden. Im nächsten Schritt wird das System under Test mit den entsprechenden Testdaten getestet. Die KI-Algorithmen werden dabei aus dem ai.store zur Verfügung gestellt. Nach der Berechnung der Metriken auf den jeweiligen Testlauf werden die Metrik-Ergebnisse in den ai.store zurückgespielt und an nLoop übertragen (siehe Abbildung 52).

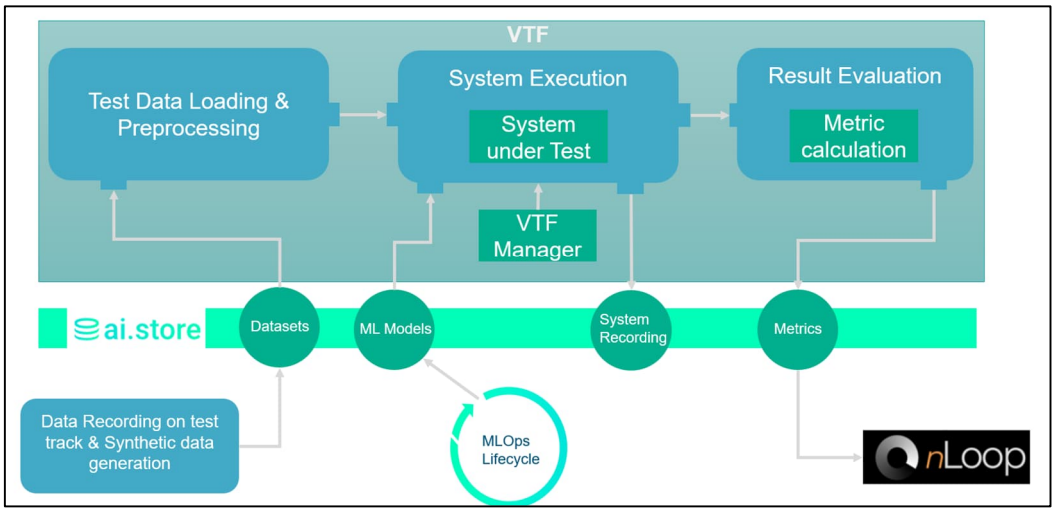


Abbildung 52: Architektur des virtuellen Testfeldes

2.1.4.4 Testszenarien und Generierung von realen und synthetischen Daten

Bei der konzeptionellen Erarbeitung der Testszenarien im Rahmen der Teststrategie wurde sich im Projekt auf das „Szenario-basierte Testen“, einem bewährten Ansatz aus dem Automobil-Sektor, gestützt. Dies ist ein systematischer Ansatz, mit dem möglichst genau die im realen Feld erwarteten Szenarien abgebildet werden können. Die Testfälle werden so definiert, dass sie sowohl die Anforderungen als auch die Aspekte der ODD berücksichtigten.

Dafür wurde die Szenario-Beschreibung in drei Abstraktionsstufen gegliedert:

- Funktionale Szenarien beinhalten sprachliche Beschreibungen, welche sich auf die ODD-Taxonomie beziehen.
- Logische Szenarien beschreiben Parameterbereiche im Zustandsraum, die aus den ODD-Attributen abgeleitet sind.
- Konkrete Szenarien stellen am Ende die Repräsentation aus dem logischen Szenario dar.

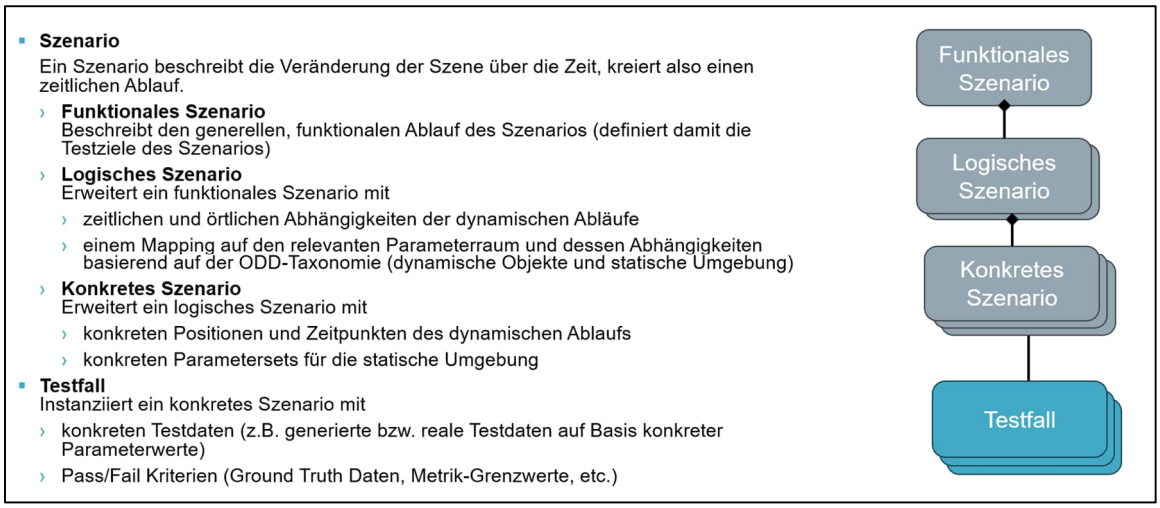


Abbildung 53: Stufen des Szenario-basierten Testens

Mithilfe dieser Methodik konnte Siemens Mobility die Testfälle für das im Projekt entwickelte System under Test definieren. Damit wurde die Basis geschaffen für die Generierung von synthetischen und realen Daten.

Als Testumgebung wurde sich in safe.trAIn auf das Gelände der Havelländische Eisenbahn Aktiengesellschaft (HVLE) in Berlin Spandau geeinigt, da Siemens Mobility dies bereits für andere Projekte zum Testen des Perzeptionssystem nutzt und dort einen realen Testträger mit Sensorik besitzt. Diese Testumgebung stellt realistische Bedingungen bereit, unter denen

verschiedene sicherheitskritische Situationen (Anfahrten auf ausgewählte Hindernisse wie z. B. Personen Dummys oder Auto Dummys) nachgestellt werden können.

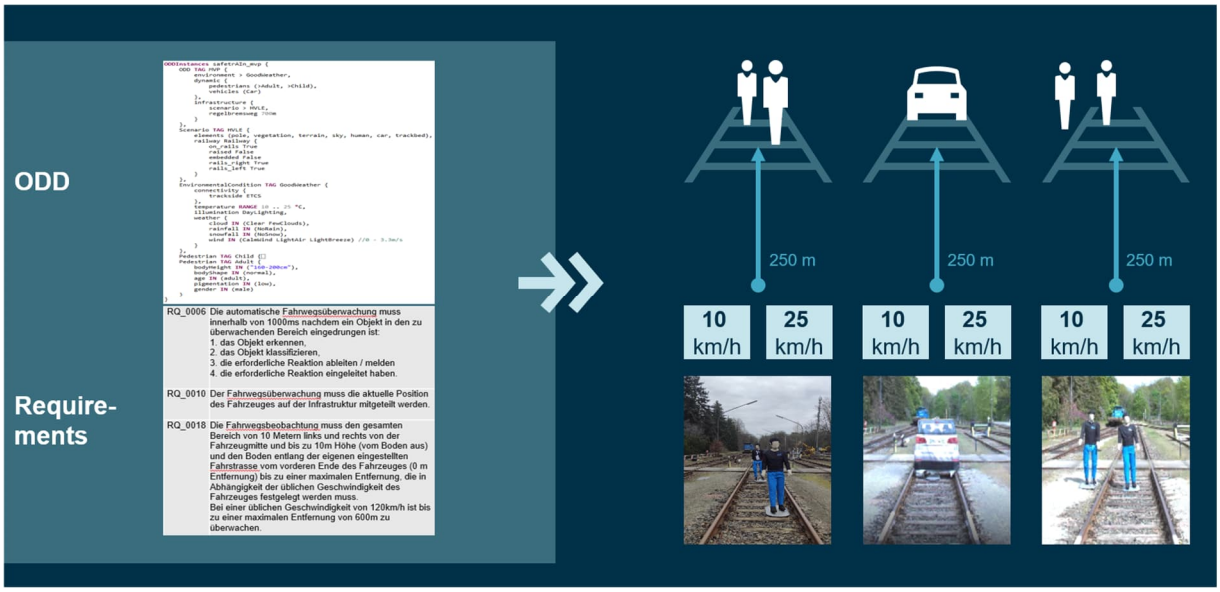


Abbildung 54: Definition von exemplarischen Testszenarien

In dem Zuge hat Siemens Mobility die nötigen realen Test-Daten für das Konsortium generiert und mittels ai.store zur Verfügung gestellt, welche in einem gemeinsamen Ansatz mit der Siemens AG und Setlabs Research GmbH entsprechend eines im Projekt entwickelten Labeling-Konzepts annotiert wurden.

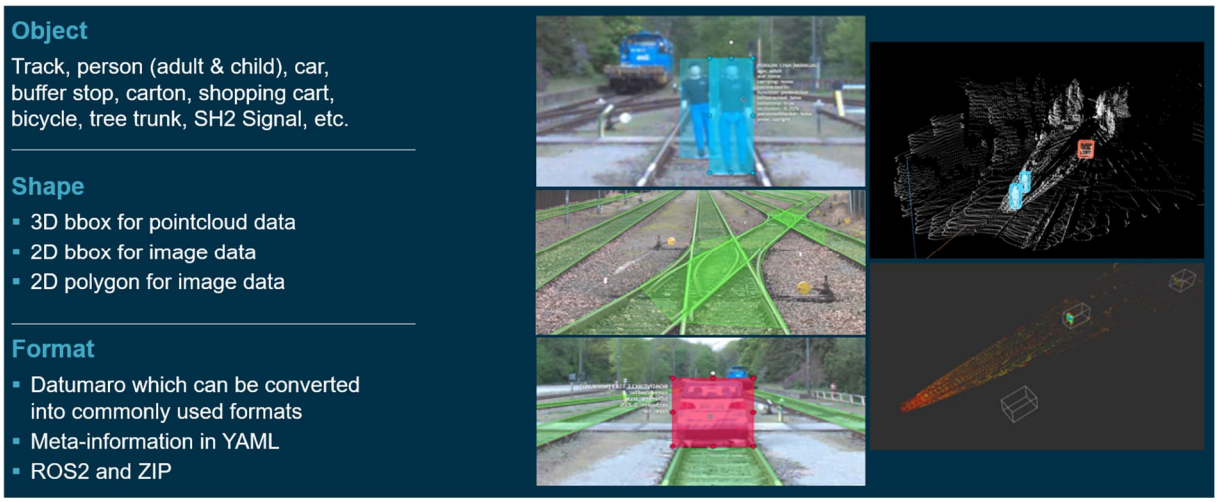


Abbildung 55: Labeling

Dieses Labeling-Konzept samt Labeling-Prozess umfasst sowohl manuelle als auch automatische Validierungsschritte, um sicherzustellen, dass die für die KI-Modelle genutzten Trainings- und Testdaten präzise und einheitlich gelabelt sind (siehe auch Kapitel 2.1.2.1).

Zur Generierung der synthetischen Testdaten durch den Projektpartner BIT Technology Solution GmbH stand Siemens Mobility im engen Austausch mit diesem und unterstützte mit relevanten Informationen, so dass neben den realen Daten auch synthetische Daten für das Virtuelle Testfeld genutzt werden konnten. Diese Kombination aus realen, auf der HVLE erfassten Daten, und synthetischen Daten stellt eine wichtige Grundlage für die Validierung der Performanz und Sicherheit des KI-basierten Perzeptionssystems dar.

2.1.4.5 Evaluierung und Beitrag zur Sicherheitsnachweisführung

Mittels der in AP2 (siehe Kapitel 2.1.2) definierten Metriken, der definierten Testszenarien und Testdaten konnten im Projekt empirische Tests der KI-Funktionen (System under Test) im virtuellen Testfeld durchgeführt werden. Diese ermöglichten es zum einen nötige Evidenzen für das Sicherheitsnachweis-Konzept zu generieren und zum anderen die Eignung der Methoden und Metriken für die Sicherheitsnachweis-Argumentation zu evaluieren. Darüber hinaus konnten basierend auf den Tests die KI-Funktionen und Methoden iterativ verbessert werden.

Alle Ergebnisse, die im Rahmen des Projektes erzeugt wurden, und deren Interpretation durch Experten wurden in Validierungs-Reports dokumentiert, um sich im Sicherheitsnachweis-Konzept darauf stützen zu können (siehe auch Kapitel 2.1.3.5). In dem Zuge entstanden ein übergeordneter Validierungs-Report, der sich auf die Ergebnisse auf Gesamtsystem-Ebene mittels der Performanz-Metriken (Precision und Recall) fokussierte, und KI-Validierungs-Reports auf Komponenten-Ebene, in dem die Ergebnisse der jeweiligen KI-Komponenten hinsichtlich der KI-Gefährdungen (aus der Landscape of AI Safety Concerns) dokumentiert und interpretiert wurden.

2.1.4.6 Beispielhafte Begutachtung

Auf Basis der gesammelten Erkenntnisse, der entwickelten Konzepte für die Sicherheitsargumentation und der generierten Testergebnisse wurde im Projekt seitens der TÜV-Organisationen ein Konzept-Gutachten erarbeitet. Dazu wurde das in AP3 entwickelte Sicherheitsnachweis-Konzept einer gesamtheitlichen gutachterlichen Bewertung unterzogen. Da das Perzeptionssystem im Rahmen des Projektes nur prototypisch in Form eines System

under Tests entwickelt wurde, befasst sich das Konzept-Gutachten damit, wie die Eignung der verschiedenen KI-basierten Funktion nachgewiesen wurden bzw. künftig nachgewiesen werden sollen. Das Konzept-Gutachten hat gezeigt, dass die entwickelten Validierungs-Ansätze und die grundsätzliche Vorgehensweise plausibel und nachvollziehbar sind, unter Vorbehalt bezüglich des aktuell prototypischen Stadiums.

Zentrale Ergebnisse beinhalten die detaillierte Ausarbeitung von spezifischen Sicherheits-Maßnahmen für KI-Funktionen und der Nachweisbarkeit von deren Wirksamkeit. Eine eindeutige Architekturdefinition des Systems bietet die Grundlage für die Sicherheitsnachweise und unterstützt die systematische Analyse möglicher Fehlermodi. Zentrale Ergebnisse beinhalten die detaillierte Ausarbeitung von spezifischen Sicherheits-Maßnahmen für KI-Funktionen und der Nachweisbarkeit von deren Wirksamkeit.

Das Projekt konnte erhebliche Fortschritte bei der Erstellung einer umfassenden Struktur für den Sicherheitsnachweis eines KI-basierten Perzeptionssystem vorweisen. Die notwendigen Dokumente zur Analyse und Sicherheitsargumentation wurden identifiziert und im Projekt systematisch erstellt. Technische Konzepte wurden entwickelt, implementiert und in realistischen Testszenarien erfolgreich validiert. Diese praktische Umsetzung ermöglichte es, Klarheit über die notwendigen Maßnahmen, Prozesse und Validierungsschritte zu schaffen, die für eine sichere Implementierung und einen zuverlässigen Betrieb der KI-basierten Perzeptionssystems erforderlich sind.

Die erzielten Ergebnisse liefern wichtige Erkenntnisse und Empfehlungen für zukünftige Projekte, insbesondere hinsichtlich der Sicherheitsvalidierung und der praktischen Anwendbarkeit von KI im Bereich vollautomatisierter Regionalzüge.

Damit bildet das Konzept-Gutachten ein Gerüst für ein zukünftiges Gutachten im Sinne der EN 50129 für Systeme, die KI-Funktionen enthalten, und setzt die Grundlage für die Zulassung eines fahrerlosen Regionalzuges mit KI-basierten Perzeptionssystem.

2.1.5 AP 5: Standardisierung und Verbreitung

Um Entwicklungen und Ergebnisse möglichst frühzeitig in die Standardisierung zu überführen und damit eine schnelle Implementierung und Akzeptanz im Markt zu erlangen, hat sich AP5 mit der Identifikation und Umsetzung von Standardisierungspotentialen, der Übertragung der Ergebnisse auf andere Anwendungsbereiche und der Ergebnisverbreitung beschäftigt (siehe dazu auch Kapitel 2.1.5).

In enger Zusammenarbeit und initiiert von DIN e. V. und VDE e. V. konnten im Projekt mehrere Standardisierungspotentiale identifiziert werden, wovon 2 Themen in eine DIN DKE SPEC gemündet sind. Die DIN DKE SPEC 99002 „Terminology: AI in railway application“ wurde von Siemens Mobility initiiert, bei zweiterer hat Siemens Mobility aktiv mitgewirkt hat.

- DIN DKE SPEC 99002 „Terminologie – KI in Bahnanwendungen“
  Diese DIN DKE SPEC legt eine einheitliche Terminologie für den Einsatz von Künstlicher Intelligenz (KI) im Bahnwesen fest. Ziel ist es, ein gemeinsames Verständnis zwischen Herstellern, Betreibern, Prüforganisationen und der Wissenschaft zu schaffen, um die Kommunikation und Zusammenarbeit zu verbessern. Die Norm definiert keine technischen Anforderungen an Algorithmen oder Prüfverfahren, sondern dient als sprachliche Grundlage für zukünftige Standards und Anwendungen.

- DIN DKE SPEC 99004 „Spezifikation von ODD im Schienenverkehr“
  Diese DIN DKE SPEC beschreibt, wie die Operational Design Domain für KI-Systeme im Schienenverkehr präzise definiert werden kann. Sie legt fest, unter welchen Bedingungen (z. B. Infrastruktur, Wetter, Verkehrsregeln) ein solches System sicher betrieben werden darf. Damit schafft sie eine wichtige Grundlage für die Entwicklung und Zulassung vollautomatisierter Züge und unterstützt internationale Standardisierungsprozesse.

Darüber hinaus hat Siemens Mobility die Erkenntnisse, insbesondere in Bezug auf die Sicherheits-Architektur auf europäischer Ebene in die Normung eingebracht.

Weiterhin fanden zu ausgewählten Themen wie z. B. der Operational Design Domain (ODD) Anwenderkreis-Meetings mit Projekt-Externen Stakeholdern statt. Neben zahlreichen Konferenz-Beiträgen und Paper-Veröffentlichungen hat Siemens Mobility in Zusammenarbeit mit der Siemens AG eine eigene Projekt-Webseite (www.safetrain-project.de) erarbeitet und

diese über den Projektverlauf aktualisiert. Auf der Innotrans24 hat Siemens Mobility die Projektergebnisse einem breiten Bahn-Fachpublikum vorgestellt.

Im Zuge der Standardisierungsarbeiten wurde sich auch mit der möglichen Verbindung der beiden Rechtsrahmen New Legislative Framwork (NLF) und Cyber Security Act (CSA) beschäftigt, dazu wurden in Form eines Workshops die relevanten Akteure zur Umsetzung einer potentiellen NLF-CSA-Brücke eingebunden. Dabei hat sich herausgestellt, dass der Cyber Resilience Act (CRA) die NLF-CSA-Brücke nicht nötig macht, da er Cyber-Sicherheitsanforderungen für Produkte und Dienstleistungen verbindlich festlegt und diese im Rahmen des NLFs durch Normen und Zertifizierungen konkretisiert. Somit wurde die Fragestellung im Projektverlauf hinfällig.

2.1.6 Zusammenfassung und Schlussgedanken

Im Kontext der Herausforderungen, die der Einsatz von KI im sicherheitskritischen Bahnumfeld für das fahrerlose Fahren mit sich bringt, und der gesetzten Projektziele konnte das safe.trAIn Projekt bedeutende Fortschritt bei der Entwicklung einer strukturierten Vorgehensweise für den Sicherheitsnachweis eines KI-basierten Perzeptionssystem erzielen.

Die dafür erforderlichen KI relevanten Sicherheitsbedenken (AI Safety Concerns) und deren Mitigations-Maßnahmen konnten identifiziert und ein Konzept zur strukturierten Sicherheitsargumentation inkl. Sicherheitsziel und den dafür notwendigen Evidenzen erarbeitet werden.

Die eindeutige Definition der Systemarchitektur bildet die Basis für den Sicherheitsnachweis und unterstützt die strukturierte Untersuchung möglicher Fehlerquellen.

Die Balance zwischen den fünf Säulen der Sicherheitsnachweis-Strategie und wie sie sich gegenseitig in ihren Schwächen ausgleichen können, gibt die Vorgabe für die Validierung eines solchen System.

Ein wichtiger Bestandteil ist auch die Entwicklung des safeMLOps Prozesses, der den traditionellen Entwicklungszyklus für Systeme (Systemengineering) mit dem spezialisierten Entwicklungszyklus für KI-Funktionen verbindet.

Um Ergebnisse möglichst frühzeitig in die Standardisierung zu überführen und damit eine schnelle Implementierung und Akzeptanz im Markt zu erlangen, wurden im Projekt zwei DIN DKE SPECS initiiert.

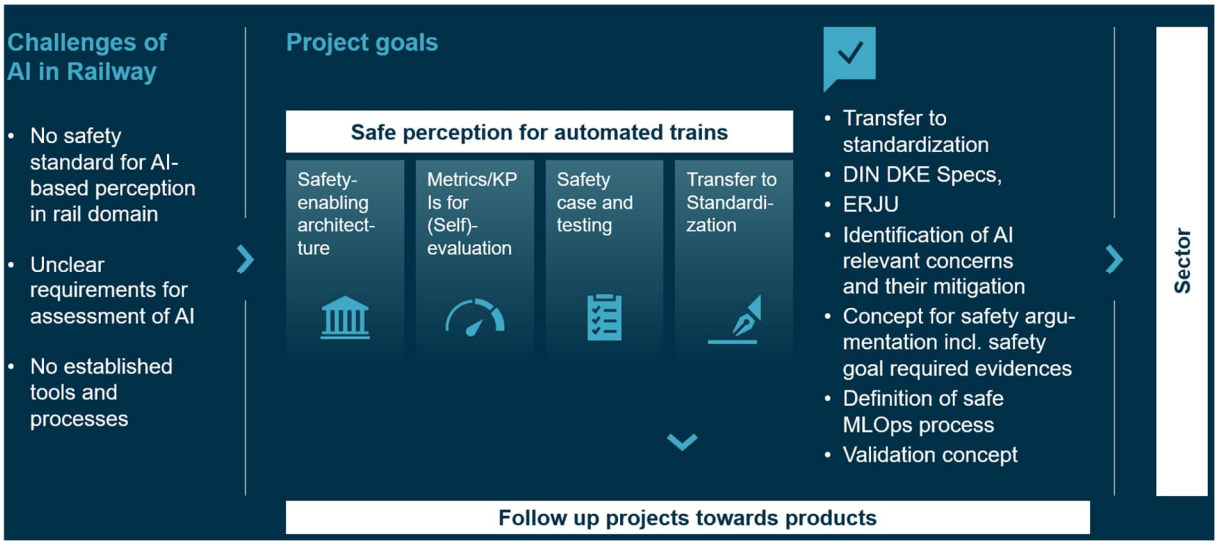


Abbildung 56: Zusammenfassende Darstellung der Projektergebnisse im Kontext der Projekt-Herausforderungen und -Ziele

Die Erstellung eines umfassenden Sicherheitsnachweises für ein KI-basiertes Perzeptionsystem eines vollautomatisierten Regionalzuges erfordert einen erheblichen Aufwand, insbesondere aufgrund der Komplexität und Unsicherheiten, die mit maschinellem Lernen verbunden sind. Entscheidend dabei ist, dass eine ausgewogene Balance zwischen verschiedenen Sicherheits-Maßnahmen und den dazugehörigen Evidenzen gefunden wird. Eine Schlüsselrolle spielt hier die Automatisierung bei der Erzeugung und Validierung dieser Evidenzen. Automatisierte Prozesse erhöhen nicht nur die Effizienz der Sicherheitsvalidierung, sondern ermöglichen auch schnelle und kontinuierliche iterative Entwicklungszyklen.

Ein wichtiger Bestandteil des Sicherheitsnachweises ist der Nachweis von Kausalitäten auf unterschiedlichen Systemebenen. Hierdurch wird sichergestellt, dass Entscheidungen der KI-Komponenten nicht nur korrekt, sondern auch nachvollziehbar sind und auf nachvollziehbaren Ursachen beruhen.

Schließlich ist die strikte Einhaltung systematischer und stringenter System-Engineering-Prinzipien unerlässlich, um sowohl die Komplexität zu bewältigen als auch den Sicherheitsnachweis nachhaltig zu unterstützen. Diese Prinzipien sorgen für Klarheit, Transparenz und Zuverlässigkeit im gesamten Sicherheitsprozess.

2.2 Wichtigste Positionen des zahlenmäßigen Nachweises

Zur Erreichung der genannten Ergebnisse plante Siemens Mobility die folgenden Kostenpositionen und setzte sie entsprechend um:

- Personalkosten: Siemens interne Mitarbeiter im zuständigen Technologiefeld
- Sonstige unmittelbare Vorhabenskosten: Cloud und Server-Kosten für die Speicherung und Verarbeitung von Trainings- und Test-Daten sowie KI-Modellen und Algorithmen
- FE-Fremdleistungen (externe Forschungs- und Entwicklungsleistungen
- weitere Kosten (wie Verwaltungs- und Reise-Kosten)

Während der Projektlaufzeit wurde jedoch deutlich, dass durch eine zunehmende Fluktuation der Mitarbeiter die internen Personalressourcen von Siemens Mobility nicht ausreichten, um die Projektziele wie geplant zu erreichen. Aus diesem Grund beantragte Siemens sowohl im Jahr 2022 als auch 2024 eine Mittelumwidmung. Diese Anpassung ermöglichte den Einsatz externer Mitarbeiter mit dem erforderlichen Fachwissen im Projekt. Durch diese strategische Neuausrichtung der Ressourcen stellte Siemens Mobility sicher, dass die erforderliche Expertise zur Verfügung stand, um den Projekterfolg trotz interner Herausforderungen zu gewährleisten.

2.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Angesichts der enormen Herausforderungen, die der Einsatz von KI im sicherheitskritischen Bahnumfeld für das fahrerlose Fahren mit sich bringt, waren die geleisteten Arbeiten im Projekt von grundlegender Bedeutung. Sie waren notwendig und angemessen, um die Basis für die nächsten Schritte in der Automatisierung des Schienenverkehrs zu legen. Dazu gehört insbesondere die Tatsache, dass es bis dato noch keine Normen, etablierten Prozesse und Vorgaben für die Entwicklung und Begutachtung eines KI-basierten Perzeptionssytems gibt. Diese Lücke stellte eine große Herausforderung dar, weshalb die Kooperation von Siemens Mobility mit verschiedenen externen Experten und Institutionen ein zentraler Bestandteil der Projektarbeit war. Diese Zusammenarbeit ermöglichte einen umfassenden Wissenstransfer zwischen unterschiedlichen Sektoren und brachte verschiedene Perspektiven ein, die wesentlich zur erfolgreichen Entwicklung im Projekt beitrugen.

Die im Rahmen des Projekts entwickelten Methoden und Ansätze legen den Grundstein für die Sicherheitsargumentation eines KI-basierten Perzeptionssystems. Diese Ergebnisse betonen die Angemessenheit der eingesetzten Mittel und der geleisteten Arbeit. In Anbetracht der Wichtigkeit und Aktualität des Themas konnte das Projekt zudem bedeutende Impulse für die gegenwärtige und zukünftige Forschung in diesem Gebiet geben.

Besonders hervorzuheben ist, dass das Projekt aufgrund seiner Relevanz für den künftigen Einsatz von KI-Systemen in sicherheitskritischen Umgebungen als offizielles Leuchtturmprojekt der „Normungsroadmap KI“ erkannt wurde. Daher hat das Projekt besondere Aufmerksamkeit bei den Stakeholdern der Normung erfahren, was seine Bedeutung und den Einfluss auf die Entwicklung zukünftiger Normen und Standards unterstreicht. Somit hat das Projekt nicht nur grundlegend zur Weiterentwicklung der Automatisierung im Schienenverkehr beigetragen, sondern auch einen entscheidenden Schritt in Richtung einheitlicher und solider Normen für KI-Perzeptionssysteme gemacht.

2.4 Voraussichtlicher Nutzen, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Als führender internationaler Anbieter von Produkten, Systemen und Lösungen, die den effizienten, sicheren und umweltfreundlichen Transport von Personen und Gütern auf der Schiene ermöglichen, hat Siemens Mobility naturgemäss ein starkes Interesse, die Automatisierungs- und Digitalisierungstechnologien für den Schienenverkehr weiterzuentwickeln. Durch die Weiterentwicklung von Automatisierungstechnologien, insbesondere mit dem Einsatz von Künstlicher Intelligenz, kann die Mobilität noch flexibler, durchsatzoptimierter, kostengünstiger, inklusiver und umweltfreundlicher gestaltet werden. Durch die gute Marktposition von Siemens Mobility im Bereich der Triebzüge sind die Voraussetzungen dafür gegeben.

Die Ergebnisse dieses Projektes stellen eine wichtige Grundlage für den sicheren und vertrauensvollen Einsatz eines KI-basierten Perzeptionssystems dar, welches die Kerntechnologie für einen zukünftigen voll-automatisierten Fahrbetrieb (GoA3/4) ist. Die im Projekt entwickelten Methoden und Konzepte steigern die Innovationsfähigkeit von Siemens Mobility, während die Risiken, Kosten und Projektlaufzeiten für zukünftige GoA3/4-Projekte reduziert werden können. Der Einsatz von KI-basierten Systemen auch in sicherheitskritischen Umgebungen, wie es bei voll-automatisierten Fahrbetrieb der Fall ist, stellt einen wesentlichen Wettbewerbsvorteil dar.

Die erzielten Ergebnisse werden bei Siemens Mobility als Grundlage für weitere Forschungsaktivitäten im Bereich des GoA3/4-Betriebes genutzt. In Folgeprojekten wie „AutomatedTrain“ konnten Aspekte wie z. B. die ODD aus safe.trAIn übertragen werden und so die Entwicklungs-Tätigkeiten von Siemens Mobility auf dem Technologiefeld voranbringen. Ebenso fließen die Erkenntnisse in die zukünftigen Forschungsaktivitäten zum ferngesteuerten Fahren von Schienenfahrzeugen (Remote Train Operation) ein.

Im Laufe des Projektes hat Siemens Mobility, teils gemeinsam mit der Siemens AG, 25 Patente zu Verfahren und Methoden eingereicht und so seine Innovationsstärke auch für den Innovations-Standort Deutschland aufgezeigt.

Die Erfolgsaussichten sind weiterhin sehr positiv. Der Druck, einen voll-automatisierten Regionalzugbetrieb zu entwickeln, ist nach wie vor sehr hoch. Dabei spielen nicht nur die wachsende Notwendigkeit der Energie- und Verkehrswende (Dekarbonisierungsziele der Bundesregierung) eine Rolle, sondern auch der zunehmende Personalmangel bei den Triebfahrzeugführern. Nur durch einen voll-automatisierten Zugbetrieb sind die erforderlichen Taktzeiten und die nötige Flexibilität möglich, um mehr Menschen zur Nutzung öffentlicher

Verkehrsmittel zu motivieren. Gerade die Eisenbahn-Verkehrs-Unternehmen stehen vor dem ernstzunehmenden Problem, dass in den nächsten Jahren mehrere Tausend Triebfahrzeugführer fehlen, weil sie in Ruhestand gehen und nicht genug Nachwuchs nachkommt.

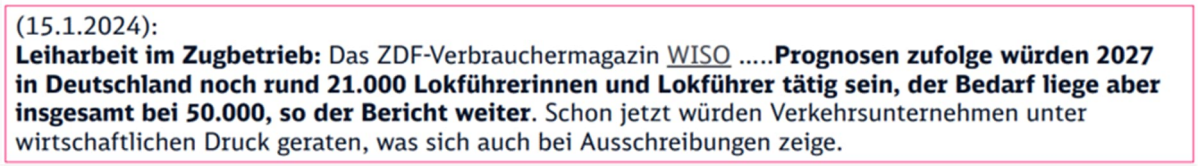

(15.1.2024):
Leiharbeit im Zugbetrieb: Das ZDF-Verbrauchermagazin WISO .....**Prognosen zufolge würden 2027 in Deutschland noch rund 21.000 Lokführerinnen und Lokführer tätig sein, der Bedarf liege aber insgesamt bei 50.000, so der Bericht weiter**. Schon jetzt würden Verkehrsunternehmen unter wirtschaftlichen Druck geraten, was sich auch bei Ausschreibungen zeige.

Abbildung 57: ZDF WISO Analyse Triebfahrzeugführer-Mangel

Dem Problem soll mittels Automatisierung entgegengewirkt werden, um den Zugbetrieb mit weniger Personal weiterhin aufrecht zu erhalten. Dies spiegelt sich auch in den gestiegenen Kundenanfragen nach automatisiertem Fahrbetrieb (GoA3/4-Betrieb) wider.

Marktanalysen und Ausschreibungen bestätigen, dass es einen wachsenden Markt für voll-automatisierte Regionalzüge gibt, wofür ein KI-basiertes Perzeptionssystem ein wesentlicher Bestandteil ist. Dies zeigt sich nicht nur an den vermehrten Kundenanfragen, sondern auch durch verstärkte Aktivitäten in dem Bereich bei den Wettbewerbern. Auf der Innotrans 2024, einer innovativen Fachmesse für Bahn- und Verkehrstechnik, zeigte sich dieser Trend zum automatisierten Bahnbetrieb bei vielen Wettbewerbern und Zulieferern.

2.5 Während der Durchführung des Vorhabens dem ZE bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Die Fortschritte auf dem Gebiet des safe.trAIn Vorhabens bei anderen Stellen, sowohl national als auch international, haben gezeigt, dass die Entwicklung voll-automatisierter Systeme für den Schienenverkehr ein dynamisches und vielfältiges Forschungsgebiet ist. Die Entwicklung eines Sicherheitsnachweis-Konzeptes für ein sicheres KI-basierte Perzeptionssystem ist dabei ein essenzieller Baustein. Sowohl auf nationaler als auch auf internationaler / europäischer Ebene gibt es Forschungsaktivitäten mit ähnlichen Entwicklungszielen.

Auf nationaler Ebene zählen die beiden Projekte „KI-Lok“ und „ARTE“ dazu. Diese sind genau wie safe.trAIn auf die Entwicklung von Automatisierungs-Technologien für den Schienenverkehr (fahrerloses Fahren in offenen Umgebungen) ausgerichtet, unterscheiden sich jedoch in ihren spezifischen Zielen und Anwendungsbereichen.

- Das „KI-Lok“ Projekt ist ein Forschungs- und Entwicklungsprojekt, das sich auf die Entwicklung von KI-basierten Steuerungssystemen für Lokomotiven konzentriert. Ziel des Projekts ist die Verbesserung der Effizienz und Sicherheit des Schienenverkehrs durch den Einsatz von KI. Während „KI-Lok“ sich auf die Entwicklung von KI-basierten Steuerungssystemen für Lokomotiven zur Optimierung der Fahrt und Verbesserung der Sicherheit fokussiert, liegt der Schwerpunkt bei safe.trAIn auf der Entwicklung eines Sicherheitsnachweis-Konzeptes für ein KI-basiertes Perzeptionssystem für einen Regionalzug.

- Das ARTE Projekt (Autonomous Rail Technology Enhancement) erforscht die technische Machbarkeit eines GoA3-Betriebs auf einer Strecke in Niedersachsen. Bei der Entwicklung der Automatisierungs-Technologien liegt der Fokus in diesem Projekt jedoch für sicherheitsrelevante Funktionen auf deterministischen Technologien weniger auf KI-Funktionen. Hier grenzt sich das safe.trAIn ab, da es sich auf die Konzipierung der Sicherheitsnachweisführung für ein KI-basiertes Perzeptionssystem fokussiert.

Im Rahmen von safe.trAIn gab es einen Austausch mit den beiden erstgenannten Projekten sowie einen vom Fördergeber veranstalteten Workshop, an dem weitere Projekte in dem Forschungsfeld teilgenommen hatten.

Das Förderprojekt „AutomatedTrain“, an dem Siemens Mobility ebenfalls beteiligt ist, befasst sich mit der technischen Machbarkeit der voll-automatisierten Bereitstellungs- und

Abstellungsfahrt. Im Vergleich zu safe.trAIn wird dabei für die Hinderniserkennung komplett auf nicht-KI Funktionalitäten gesetzt. Dennoch erfolgte eine enge Abstimmung zwischen beiden Projekten, insbesondere hinsichtlich Anforderungen, Architektur und ODD-Methodik, um die Erkenntnisse zur möglichen Zulassbarkeit eines GoA4-Systems und von KI zu synchronisieren.

Ein weitere Forschungsaktivität auf dem Gebiet ist das französische Forschungsprogramm Confiance.ai, das im Rahmen der Strategie „France 2030“ gefördert wurde. Es hat zum Ziel, vertrauenswürdige KI (Trustworthy AI) für kritische industrielle Systeme zu ermöglichen. Anders als safe.trAIn ist Confiance.ai branchenübergreifend: In einem Ökosystem von über 50 Partnern – große Industrieunternehmen (Airbus, Renault, Thales, Valeo, u.a.), Forschungsinstitute (CEA, Inria, IRTs) und Start-ups – wurden *Methoden und ein Software-Toolkit* entwickelt, mit dem Hersteller KI in sicherheitskritische Produkte integrieren können. Confiance.ai adressiert mehrere Anwendungsfälle aus unterschiedlichen Domänen, etwa visuelle Qualitätskontrolle in der Fertigung (Schweißnahtprüfung bei Renault), Objekterkennung für autonome Fahrzeuge (Valeo), Kollisionsvermeidung in unbemannten Flugsystemen (Airbus), Nachfrageprognosen (Air Liquide) usw.. Das Programm legte seinen Schwerpunkt auf eine ganzheitliche ingenieurmäßige Methode (ein W-Modell als Entwicklungslebenszyklus für KI) sowie auf ein Portfolio von Tools, um Anforderungen wie Robustheit, Erklärbarkeit, Datengüte, etc. entlang des KI-Lebenszyklus sicherzustellen. Confiance.ai versteht sich als „vertrauenswürdige KI-Community“, die über das Programm hinaus fortbesteht und die entwickelten Ergebnisse in die Breite trägt. Es gilt als ein Leuchtturm der französischen KI-Strategie und wurde eng mit der Entwicklung europäischer KI-Regulierungen (AI Act) verzahnt. Der Unterschied zu safe.trAIn liegt darin, dass in Confiance.ai ein branchenübergreifenden Ansatz verfolgt wird: Die Domäne ist nicht auf einen Sektor beschränkt, sondern umfasst verschiedene sicherheitskritische Industrien, von Automotive über Luftfahrt bis Fertigung. Die Zielsetzungen überschneiden sich insofern, als beide Programme vertrauenswürdige KI-Technologien für kritische Anwendungen etablieren wollen. Doch safe.trAIn tut dies exemplarisch am Use-Case eines fahrerlosen Regionalzug, während Confiance.ai eine generische Methodik für vertrauenswürdige KI in beliebigen Missionskritischen Systemen anstrebt.

2.6 Erfolgte oder geplante Veröffentlichungen des Ergebnisses nach Nr. 11

Im Zuge des Projekts hat Siemens Mobility die gewonnenen Erkenntnisse und Ergebnisse auf einer Vielzahl von Konferenzen und wissenschaftlichen Veranstaltungen sowohl im Bahnsektor als auch in der KI-Community umfassend präsentiert.

Nachfolgend die wichtigsten Konferenzen aufgelistet:

- AITA Symposium im Rahmen der AAAI-Konferenz 2023
- AITA 2023
- 32. safeTRANS Industrial Day 2023
- IOT@Siemens 2023
- Confiance.AI Day 2024
- Innotrans 2024
- CVPR Workshop SAIAD2024 (Safe AI in All Domains)
- Digital Rail Summer School 2025
- TÜV Süd safe.tech Konferenz 2025
- CVPR Workshop SAIAD2025 (Safe AI in All Domains)
- ATRASS Series – Responsible AI https://www.youtube.com/watch?v=Q_9ByhG9grI&t=1507s)
- Video-Interview zu safe.trAIn im Zuge der Siemens Mobility AI Compaign https://www.linkedin.com/posts/siemens-mobility_transformmobilityforeveryone-activity-7336357605709176834-Kgj1?utm_source=share&utm_medium=member_desktop&rcm=ACoAACftXpIBtxd2Xy2aNbYyHbzzF_kx4LSFL7g

Diese Präsentationen dienten dazu, die Projektergebnisse nicht nur bei Experten und Fachleuten bekannt zu machen, sondern auch die breite Diskussion über innovative Ansätze zur Nutzung von KI im Bereich des Schienenverkehrs zu fördern. Dabei wurden auch zu verschiedenen Themen umfangreiche technische Papers veröffentlicht, die die detaillierten Forschungsergebnisse und technologischen Fortschritte dokumentieren.

Ein bedeutender Aspekt der Wissensweitergabe bestand darin, einige Ergebnisse in die wissenschaftliche Lehre zu integrieren. So flossen Erkenntnisse aus dem Projekt in die Vorlesungen „Autonomes Fahren“ und „Trustworthy Machine Learning Systems“ an der Technischen Universität München ein. Dies bereitet die heranwachsende Generation von Ingenieuren und Wissenschaftlern auf zukünftige Herausforderungen im Bereich der Künstlichen Intelligenz und des vollautomatisierten Fahrens vor.

Auf der Innotrans 2024, einer führenden Fachmesse für Bahn- und Verkehrstechnik, präsentierte Siemens Mobility das Projekt an einer eigens eingerichteten Dialogstation. Diese Plattform bot einem breiten Publikum die Möglichkeit, sich direkt über die gewonnenen Erkenntnisse zu informieren.

Zudem engagierte sich Siemens Mobility bei der Erarbeitung der beiden DIN DKE SPECs:

- DIN DKE SPEC 99002 „Terminologie – KI in Bahnanwendungen“
- DIN DKE SPEC 99004 „Spezifikation von ODD im Schienenverkehr“

Bei der DIN DKE SPEC 99002 „Terminologie – KI in Bahnanwendungen“ hat Siemens Mobility als Initiator gewirkt.

Darüber hinaus wurde das Wissen aus dem Projekt, insbesondere in Bezug auf Sicherheitsarchitekturen, auf europäischer Ebene in Normungsprozesse eingebracht und somit maßgeblich zur Entwicklung standardisierter Verfahren beigetragen.

3 Abkürzungsverzeichnis und Referenzen

Abkürzungen:

ASAM	Association for Standardization of Automations and Measuring Systems
AStriD	Autonomen Straßenbahn im Depot
ATO	Automatic Train Operation
BerDiBa	Berliner Digitaler Bahnbetrieb
CSM-RA	Common Safety Method for Risk Evaluation and Assessment
DIN	Deutsches Institut für Normung
DZSF	Deutsche Zentrum für Schienenverkehrsforschung
EBA	Eisenbahnbundesamt
ECS	Equivalence class Sets
ECS_EE	Input space Equivalent Output space Equivalent
ECS_EU	Input space Equivalent Output space Unequivalent
ECS_UE	Input space Unequivalent Output space Equivalent
ECS_UU	Input space Unequivalent Output space Unequivalent
ERJU	Europe's Rail Joint Undertaking
ETCS	European Train Control System
ECS_EU	Input space Equivalent Output space Unequivalent
FMEA	Fehlermöglichkeits- und Einflussanalyse
GoA	Grade of Automation
HVLE	Havelländische Eisenbahn AG
JDL	Joint Directors of Laboratories
JTC21	CEN and CENELEC Joint Technical Committee 21
LAISC	Landscape of AI Safety Concerns
ML	Machine Learning
MLOps	Machine Learning Operations
MVP	Minimal Viable Product
ODD	Operational Design Domain
OoD	Out of Distribution
PFD	Probability of Failure on Demand
QI²	Quality Indicator Squared
R2DATO	Rail to Digital Automated up to Autonomous Train Operation

SHLQI²	Skaliertes Histogramm Lokaler QI²
SIL	Sicherheitsintegritätslevel
SMO	Siemens Mobility GmbH
SOTIF	Safety of the Intended Functionality
SuT	System under Test
Tf	Triebfahrzeugführer:innen
TRL	Technologiereifegrad
VDE	Verband der Elektrotechnik Elektronik Informationstechnik
VTF	Virtuelles Testfeld

Referenzen:

Steinberg e.a., 1998 Revising the JDL model, Proceedings of the SPIE

Hinzen, A. 1993, Der Einfluss des menschlichen Fehlers auf die Sicherheit der Eisenbahn, Aachen, Rhein.-Westfäl. Technische Hochschule Aachen

Walber - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36926283

ISO/IEC WD 5259-2:202X(X) Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures

Deutsche Bahn AG, "Richtlinie 408.2341: Anforderungen an Triebfahrzeugführer hinsichtlich Streckenbeobachtung," DB Netz AG, Frankfurt am Main, Germany, 2021

Schnitzer, R.; Kilian, L.; Roessner, S.; Theodorou, K.; & Zillner, S. (2024), "Landscape of AI Safety Concerns: A Methodology to Support Safety Assurance for AI-based Autonomous Systems," 8th International Conference on System Reliability and Safety (ICSRS), 2024.

Sicherichs, C.; Geerkens S.; Braun, A.; Waschulzik, T. (2022) “ECS - an Interactive Tool for Data Quality Assurance”

Sicherichs, C.; Geerkens S.; Braun, A.; Waschulzik, T. (2022) “QI2 - an Interactive Tool for Data Quality Assurance”