

Berichtsblatt

1. ISBN oder ISSN geplant	2. Berichtsart (Schlussbericht oder Veröffentlichung) Partnerspezifischer Schlussbericht (SB)
3. Titel safe.trAln – sichere Ki am Beispiel fahrerloser Regionalzug	
4. Autor(en) [Name(n), Vorname(n)] Geerkens, Simon Sieberichs, Christian	5. Abschlussdatum des Vorhabens 31.03.2025
	6. Veröffentlichungsdatum
	7. Form der Publikation Abschlussbericht
8. Durchführende Institution(en) (Name, Adresse) Hochschule Düsseldorf Münsterstr. 156 40476 Düsseldorf	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen 19I21039B
	11. Seitenzahl 13
12. Fördernde Institution (Name, Adresse) Bundesministerium für Wirtschaft und Klimaschutz (BMWK) 53107 Bonn	13. Literaturangaben 25
	14. Tabellen 0
	15. Abbildungen 0
16. Zusätzliche Angaben -	
17. Vorgelegt bei (Titel, Ort, Datum)	

18. Kurzfassung

Mit dem EU AI Act wurden KI-Systeme mit sicherheitskritischen Funktionen als Hochrisiko-Anwendungen eingestuft. Sie müssen hohen Anforderungen an Datenqualität, Robustheit, Transparenz und Konformitätsbewertung genügen. Forschungsarbeiten zur Erklärbarkeit neuronaler Netze (z. B. LRP, CAM) und zur Auswirkung optischer Eigenschaften auf KI-Systeme bilden die Basis. Zudem wurde erkannt, dass fehlerhafte Daten und optische Abbildungsfehler die Sicherheit stark beeinträchtigen können.

Ziel des Projekts *safe.train* war es, die Voraussetzungen für einen sicheren vollautomatisierten Zugbetrieb zu schaffen. Im Fokus stand die Absicherung KI-basierter Verfahren zur Hinderniserkennung und deren Integration in Sicherheitsnachweise. Die Hochschule Düsseldorf (HSD) übernahm zentrale Aufgaben bei der Entwicklung und Weiterentwicklung von Methoden zur Daten- und KI-Absicherung, insbesondere den *Queens-Methoden* (QI² und ECS).

Methoden:

- Entwicklung und Anwendung der Queens-Methoden zur Bewertung von Datenqualität und Modellrobustheit
- Analyse von Unsicherheiten in KI-gestützter Objekterkennung und deren Propagation
- Entwicklung physikalisch realistischer Optikmodelle zur Untersuchung sensorischer Einflüsse
- Validierung im virtuellen Testfeld (GoA4-System, driverless operation)
- Integration in Sicherheitsarchitekturen (RAMS- und GSN-Methodik) sowie Begutachtung durch TÜV.

Ergebnisse:

- Erfolgreiche Weiterentwicklung der Queens-Methoden (QI², ECS) für Daten- und Modellbewertung
- Entwicklung von Verfahren zur Erklärbarkeit (Kombination QI² mit LRP)
- Neue Ansätze zur datenqualitätsbasierten Fehlererkennung ohne Training
- Optische Modellierungen und verbesserte Kamerakalibrierung
- Beitrag zur Standardisierung (DIN DKE SPEC 99002)
- Mehrere wissenschaftliche Publikationen und Betreuung von Promotionen und Abschlussarbeiten.

Die entwickelten Methoden ermöglichen eine systematische Absicherung von KI-Systemen im Bahnumfeld.

Anwendungsmöglichkeiten bestehen in der Integration in Zulassungsprozesse, bei der Datenkuratierung, in der Sensorikgestaltung sowie in weiteren sicherheitskritischen Bereichen. Darüber hinaus sind die Methoden anschlussfähig für andere Forschungsfelder (z. B. Transformer-Netzwerke, chaostheoretische Analysen). Damit wurde eine technologische Grundlage für den sicheren Betrieb fahrerloser Regionalzüge geschaffen

19. Schlagwörter

Automatisierter Zugbetrieb, Fahrerloser Regionalzug, Künstliche Intelligenz (KI), Hochrisiko-KI, KI-Absicherung, Sicherheitsnachweisführung, Datenqualität, Daten-Shift, Label-Fehler, Robustheit von KI-Systemen, Erklärbarkeit / Transparenz neuronaler Netze, Queens-Methoden (QI², ECS), Edge Cases, RBF-Klassifikator, Optische Modellierung, Kamerasensorik / Kalibrierung, Sensorfusion

20. Verlag

21. Preis

Abschlussbericht

Des Projektes **safe.trAI**n - sichere Ki am Beispiel fahrerloser Regionalzug

Partner: **Hochschule Düsseldorf**



**Finanziert von der
Europäischen Union**

NextGenerationEU

Gefördert durch:



Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

Teil I:

1. Aufgabenstellung

Das Projekt safe.trAln adressierte die Programmsäule *Automatisiertes Fahren* des Förderprogramms im Bereich Schienenverkehr und verfolgte das Ziel, die Voraussetzungen für einen vollautomatisierten Zugbetrieb zu schaffen. Im Mittelpunkt stand dabei die sichere Anwendung von KI-basierten Verfahren, insbesondere im Hinblick auf das zuverlässige Erkennen von Hindernissen. Um die Möglichkeiten von Künstlicher Intelligenz mit den Sicherheitsanforderungen des Schienenverkehrs zu verbinden, wurden Methoden zur Qualitätssicherung sowie Prüf- und Validierungsverfahren für KI-gestützte Perzeptionssysteme erforscht und in eine Sicherheitsnachweisführung eingebettet. Darauf aufbauend wurde eine Sicherheitsarchitektur am Beispiel eines fahrerlosen Regionalzugs konkretisiert und ein GoA4-System (Driverless Operation) konzeptionell in einem virtuellen Testfeld entwickelt und validiert. Damit leistete safe.trAln einen wesentlichen Beitrag zur Integration von KI in die Zulassungsprozesse des Bahnumfelds und schuf die technologische Basis für den sicheren Betrieb vollautomatisierter Schienenfahrzeuge.

Im Rahmen des Projektes übernahm die HSD zentrale Aufgaben im Bereich der KI-Absicherung. Der Schwerpunkt lag in der Entwicklung und Weiterentwicklung der Queens-Methoden (QI² (integrierter Qualitätsindikator) und ECS (equivalent class sets)), die als Werkzeuge zur Untersuchung von Datenqualität und zur Analyse komplexer KI-Verfahren eingesetzt wurden. Darüber hinaus wurden die Unsicherheiten in der KI-gestützten Objekterkennung und Segmentierung untersucht und deren Propagation durch die nachfolgenden Netzwerke analysiert. Ein weiterer wesentlicher Beitrag bestand in der Komplexitätsreduktion durch optische Merkmale, deren Wirkung mit Hilfe der Queens-Methoden überprüft werden konnte. Ergänzend wurden die Queens-Methoden zur Identifikation von Edge Cases herangezogen und lieferten damit wichtige Beiträge für die Robustheit der eingesetzten ML-Methoden und Datensätze. Zudem unterstützte die HSD die Partner bei der Entwicklung und Überprüfung von RBF-Netzwerken und stellte die Queens-Methoden dem Konsortium zur Verfügung. Insgesamt wurden die Arbeiten in diesem Bereich über die ursprünglichen Ziele hinaus erfolgreich realisiert.

Ein weiterer Schwerpunkt der Arbeiten der HSD lag in der optischen Modellierung und Sensorik. Dazu wurden physikalisch-realistische Optikmodelle für Kamera-Sensoren entwickelt und angepasst, um den Einfluss optischer Eigenschaften auf Bildqualität, Sensorfusion und KI-Modelle systematisch untersuchen zu können. Ergänzend erfolgte die Analyse optischer Unsicherheiten, insbesondere im Hinblick auf Mess- und Kalibrierdrift, und deren Auswirkungen auf die Robustheit der KI-Methoden. Darüber hinaus wirkte die HSD an der Definition und Spezifikation der optischen Sensorik mit und leistete damit einen Beitrag zur konzeptionellen Ausgestaltung des Sensorsystems. Anstelle des ursprünglich geplanten Prüfstands wurden die extrinsischen und intrinsischen Kameraparameter optimiert, wodurch eine alternative Möglichkeit zur Verbesserung der Modellgüte geschaffen wurde. Trotz dieser Anpassung konnte die HSD im Bereich der optischen Modellierung substantielle Ergebnisse erzielen, die als Grundlage für die weiteren Arbeiten im Projekt dienen.

Im Bereich der Integration in die Sicherheitsarchitektur nutzte die HSD die entwickelten Optik-Modelle und Queens-Methoden, um die Sicherheitsnachweisführung zu unterstützen. Dabei wurde der Einfluss optischer Eigenschaften auf RAMS-Anforderungen sowie nicht-funktionale Anforderungen untersucht. Zudem leistete die HSD Beiträge zur Architektur des sicheren Sensorsystems, indem optische Aspekte systematisch in die Sicherheitsargumentation eingebettet wurden.

Darüber hinaus engagierte sich die HSD in der wissenschaftlichen Verwertung der Projektergebnisse. Dies umfasste die Betreuung von kooperativen Promotionen sowie die Veröffentlichung und Präsentation der erzielten Ergebnisse in wissenschaftlichen Publikationen, wodurch die Projekterkenntnisse nachhaltig in die Forschungsgemeinschaft eingebracht wurden.

2. Voraussetzungen

Für die erfolgreiche Durchführung des Vorhabens mussten zunächst wesentliche infrastrukturelle und organisatorische Voraussetzungen geschaffen werden. Hierzu gehörte insbesondere die Anschaffung leistungsfähiger Server, die als zentrale Rechenressource für die Entwicklung und Ausführung der Methoden an der HSD dienten. Ergänzend wurden für ein erstes Rapid Prototyping hardwarestarke Arbeitslaptops bereitgestellt, um eine flexible und effiziente Umsetzung einzelner Arbeitspakete lokal und schnell zu ermöglichen.

Eine weitere Voraussetzung für die erfolgreiche Durchführung des Vorhabens war die Verfügbarkeit von qualifiziertem Personal. Die zeitnahe Bereitstellung dieses wurde durch die geringe Zeit zwischen der offiziellen Projektzusage und dem Projektstart erschwert. Während für den Projektstart Expertise für die Themengebiete der KI und der Schnittstelle Optik-KI zur Verfügung stand, war dies für die reinen Optikthemen zunächst nicht gegeben. Der für diesen Themenkomplex vorgesehene Mitarbeiter konnte erst mit einer mehrmonatigen Verzögerung eingebunden werden.

Eine besondere Herausforderung stellte der Zugang zu gesicherten und gelabelten Fahrdaten dar. Diese standen nur eingeschränkt und zeitlich verzögert zur Verfügung, was die Nutzung für bestimmte Analyseaufgaben erschwerte. Hierbei sind die beiden Schwerpunkte Optik (Abgleich der Ausgaben mit Labeln bei einer rein deterministischen Fahrwegserkennung, sowie Prüfung der extrinsischen Kamerakalibrierung) und KI (ausgiebige Überprüfung der Datenqualität, sowie hinreichende Sicherheitsprüfung der KI Methodik) betroffen. Für die methodische Umsetzung wurden zudem neuronale Netze benötigt, die als Referenzbasis für die Anwendung und Validierung der im Projekt entwickelten Verfahren dienten. Zudem war die Möglichkeit, Open-Source-Software einzubringen, was insbesondere in der Implementierung der entwickelten Methoden eine zentrale Rolle spielte, eine Voraussetzung.

Auf organisatorischer Ebene war der Abschluss eines IP-Vertrags zwischen den Projektteilnehmern und der Hochschule Düsseldorf erforderlich, um gemeinsame Erfindungen rechtlich abzusichern.

Darüber hinaus mussten die geltenden Normen und Standards im Bereich Sicherheitsnachweisführung berücksichtigt werden. Diese wurden im Projekt (von anderen Stellen) systematisch erforscht, gebündelt und nach und nach in die methodische Entwicklung integriert. Ergänzend wurden spezifische Anforderungen an Sicherheitsaspekte – sowohl für KI-basierte Methoden als auch im Bereich optischer Systeme – in einem eigenen Arbeitspaket (AP3.11) adressiert und innerhalb einer Sicherheitsnachweisführung weiterentwickelt.

3. Planung und Ablauf des Vorhabens

Das Vorhaben wurde von Beginn an mit einer klaren Struktur und personeller Planung umgesetzt. Geplant war die Besetzung von drei Promotionsstellen sowie die Unterstützung durch ein bis zwei weitere wissenschaftliche Hilfskräfte. Zwei der Promotionsstellen fokussierten auf Fragestellungen der Künstlichen Intelligenz, der Datenqualität und der Schnittstelle zur Optik, während eine weitere Promotionsstelle Themen der Sensorik bearbeitete. Ergänzt wurden die Arbeiten kontinuierlich durch Bachelor- und Masterarbeiten, sodass eine enge Verzahnung von Forschung und Ausbildung gewährleistet war. Die wissenschaftliche Expertise im Bereich KI und Optik bildete dabei die tragende Grundlage für die inhaltliche Arbeit.

Inhaltlich standen die Entwicklung und Weiterentwicklung neuer Methoden sowie deren konkrete Umsetzung Vordergrund. Die Methoden sollten im Projektverlauf erforscht, angewendet und systematisch bewertet werden. Die erzielten Ergebnisse sollten dann in regelmäßigen Abstimmungen besprochen und anderen zur Verfügung gestellt werden.

Die Planung gliederte sich in mehrere Hauptstränge: (1) die Identifikation des Einflusses optischer Systeme auf KI-Verfahren, (2) die Auslegung einer geeigneten Optik, (3) die Absicherung von KI über Methoden der Erklärbarkeit und Transparenz, (4) die Absicherung der für das Training verwendeten Daten, (5) die Entwicklung eines geeigneten Sicherheitsnachweiskonzeptes für KI Systeme mithilfe der erforschten Methoden und (6) die Mitgestaltung eines sicheren MLOps Prozesses.

Diese Stränge wurden während des Vorhabens iterativ bearbeitet und in enger Abstimmung mit den Projektpartnern weiterentwickelt. Der Ablauf erforderte zunächst eine intensive Recherche und Einarbeitung in den aktuellen wissenschaftlichen und technischen Stand auf den beiden Gebieten (KI und Optik). Anschließend wurden zu entwickelnde Methoden und Verfahren zunächst grob und in einer recht theoretischen Grundlagenforschung behandelt. Mit der Mitarbeit und Weiterentwicklung der Sicherheitsaspekte, dem Sicherheitskonzept und der Sicherheitsnachweisführung ließen sich die Arbeiten konkretisieren und zielgerichtet durchführen. Vor allem die Safety Requirements, eine Einordnung der entwickelten Methoden in diese und die Operational Design Domain waren hierbei federführend. Somit wurde aus einer anfänglichen Grundlagenforschung eine konkrete Forschung und Mitarbeit an Konzepten zur sicherheitsgerichteten, anwendungsorientierten und systematischen Absicherung eines autonomen Zuges.

Am Ende stand vor allem die Zielsetzung, die entwickelten KI-Absicherungsmethoden (Datenqualität, Erklärbarkeit und Transparenz) einer sicherheitsgerichteten Bewertung zu unterziehen. Diese Bewertung erfolgte durch Gutachter des TÜV, wodurch eine unabhängige und praxisorientierte Überprüfung gewährleistet werden konnte. Ziel war es, exemplarische Absicherungen aller möglichen Aspekte im Sicherheitsnachweisconcept durchzuführen.

4. Wissenschaftlicher und technischer Stand

Mit dem EU AI Act (Verordnung (EU) 2024/1689) [1] wurde erstmals eine verbindliche Einordnung von KI-Systemen vorgenommen. Systeme, die sicherheitskritische Funktionen im Fahrzeug übernehmen, gelten dabei als „Hochrisiko-KI“ und müssen strengen Anforderungen an Datenqualität, Robustheit, Transparenz und Konformitätsbewertung genügen. Damit war bereits vor Projektbeginn klar, dass Methoden zur Absicherung und Nachweisführung einen wesentlichen Teil des Forschungsbedarfs darstellen würden. Ähnliche Schwerpunkte betont auch die Übersicht in [2], die typische AI Safety Concerns – wie Daten-Shift, Label-Fehler, mangelnde Erklärbarkeit oder Unsicherheitskalibrierung – benennt und mögliche Abmilderungsstrategien systematisch aufbereitet.

Bereits Ende der 1990er Jahre wurden mit der QUEEN-Methodik [3] (Qualitätsgesicherte effiziente Entwicklung vorwärtsgerichteter neuronaler Netze mit überwachtem Lernen) erste strukturierte Vorgehensweisen für die Entwicklung und Validierung von KI-basierten Verfahren vorgeschlagen. QUEEN verband dokumentations- und testbasierte Qualitätssicherung mit einem Entwicklungsmodell, das bereits zentrale Elemente heutiger „ML-Engineering“-Ansätze vorwegnahm. Diese Methodik und die Methoden QI^2 und ECS innerhalb von QUEEN dienen als fundamentale Basis für die Weiterentwicklung von Methoden zur KI Absicherung.

Ein weiterer Forschungsstrang, der vor allem als Orientierung und Referenz für Sicherheitsaspekte gilt, betrifft die Erklärbarkeit neuronaler Netze. Methoden wie Layer-wise Relevance Propagation (LRP) [4], DeConv-Nets [5] oder Class Activation Maps (CAM) [6], [7] haben sich als Standardverfahren etabliert, um die Entscheidungsgrundlagen komplexer Modelle sichtbar zu machen. Diese Verfahren dienen nicht nur der Plausibilisierung, sondern sind auch Bausteine für die sicherheitsgerichtete Argumentation im Sinne von Transparenz und Nachvollziehbarkeit.

Zunehmend in den Fokus rückte auch die Wechselwirkung zwischen physikalischer Optik und der Leistungsfähigkeit KI-basierter Bildverarbeitung. Arbeiten wie [8], [9], [10], [11] zeigen, dass optische Aberrationen – wie Koma, Astigmatismus oder Unschärfen durch Windschutzscheiben – einen erheblichen Einfluss auf Klassifikations- und Detektionsleistung haben können. Mit „OpticsBench“ wurde ein Benchmark entwickelt, der die Robustheit gängiger Netze gegenüber realistischen Abbildungsfehlern quantifiziert. Parallel dazu wurde die sicherheitsrelevante Dimension dieser Effekte hervorgehoben und Konzepte wie physik-informierte Kalibrierung vorgeschlagen, um Unsicherheiten bei gestörter Bildqualität besser zu erfassen.

Eng damit verknüpft ist die Frage nach der Qualität der Trainingsdaten. Studien wie [12]. haben gezeigt, dass fehlerhafte Labels und unzureichend kuratierte Datensätze zu erheblichen

Verzerrungen in der Leistungsbewertung führen. Für den sicherheitsgerichteten Einsatz von KI sind daher Strategien zur Sicherung der Datenqualität, zur Erkennung fehlerhafter Annotationen und zur Kontrolle von Daten-Shift unverzichtbar. Dies wird auch durch aktuelle Normungsaktivitäten wie ISO/TR 4804 (V&V für automatisierte Fahrsysteme) und ISO/PAS 8800:2024 (KI-Sicherheit in Straßenfahrzeugen) unterstrichen.

5. Zusammenarbeit mit anderen Stellen

In Zusammenarbeit mit SAG unterstützte die HSD in der Erstellung des Sicherheitsnachweises in ihrer Fachexpertise der KI und Optik. Dies umfasste die Mitarbeit an der ‚Landscape of safety concerns‘ in den relevanten Themengebieten, die Ausarbeitung der Validierung von Methoden und Metriken sowie deren Vorstellung und Verteidigung vor den Partnern des TÜV. Gemeinsam mit Fraunhofer IKS arbeitete die HSD zudem an der Entwicklung von GSN-Bäumen (Goal Structuring Notation) in welchen sicherheitsrelevante Einzelpunkte wie Datenqualität, Transparenz, Performance, Robustheit und die CLOCS-Fusion systematisch aufgegliedert und mit geeigneten Absicherungsstrategien und Metriken verknüpft wurden. Auf diese Weise trug die HSD wesentlich dazu bei, die entwickelten Konzepte in die sicherheitsrelevanten Nachweisprozesse zu integrieren und stand bei fachlichen Fragen zur Absicherung beratend zur Seite.

Im Bereich der optischen Sensorik arbeitete die HSD end mit Siemens Mobility zusammen. Diese Kooperation umfasste die Auswahl und Auslegung geeigneter Kamerasysteme sowie die Planung der Sensorik im Gesamtsystem. Darüber hinaus beteiligte sich die HSD an der Erhebung von Testdaten auf dem HVLE-Testgelände, welche als Grundlage für die Bewertung und Weiterentwicklung der entwickelten Methoden diente.

Eine enge Zusammenarbeit bestand zudem mit ITQ bei der Entwicklung und Absicherung eines RBF-Klassifikators. Hierfür wurden geprüfte Datensätze mit den Queens-Methoden weiter untersucht, um Stärken und Schwächen des Klassifikators systematisch herauszuarbeiten. Ergänzend wurde in Kooperation mit dem externen Dienstleister Lutz Schäfer die Nutzung von Disparitätskarten aus Stereo-Kamerasystemen erforscht, um unzuverlässige Bildbereiche zu identifizieren. Diese Informationen zur Objektdistanz konnten anschließend in den RBF-Klassifikator eingebracht werden.

Darüber hinaus beteiligte sich die HSD an der Ausarbeitung von Standards, sofern diese eine inhaltliche Überschneidung mit den eigenen Themengebieten aufwiesen. Konkret erfolgte ein Beitrag zur Erarbeitung des DIN DKE SPEC 99002 „Terminology: AI in Railway Applications“, bei dem relevante Begriffe und Definitionen im Kontext von KI im Bahnbereich abgestimmt und präzisiert wurden.

Teil II:

1. Verwendung der Zuwendung und erzielttes Ergebnis

Verwendung der Zuwendung

Die Zuwendung wurde überwiegend für Personalaufwendungen eingesetzt. 719,995.37 € entfielen auf wissenschaftliches Personal, insbesondere Promotionsstellen sowie wissenschaftliche Hilfskräfte.

- Promotionsstellen:
 - Zwei Doktorandenstellen über jeweils 3 Jahre und 3 Monate,
 - eine Doktorandenstelle über 2 Jahre und 7 Monate,
 - eine Doktorandenstelle über 6 Monate.
- Wissenschaftliche Hilfskräfte:

Drei studentische bzw. wissenschaftliche Hilfskräfte unterstützten die Arbeiten über Zeiträume von 1 Jahr und 5 Monaten, 1 Jahr und 1 Monat sowie 5 Monaten (Kostenstellen 0812, 0817, 0822).

Zur Ausstattung der Promovierenden wurden Arbeitsrechner sowie benötigte Hardware (z. B. Grafikkarten, SSDs) angeschafft (Kostenstellen 0863, 0850).

Weitere Mittel wurden für Reisekosten eingesetzt (Kostenstelle 0846). Hierunter fielen insbesondere Konsortialtreffen, Fachkonferenzen sowie Dienstreisen zur Präsentation der Projektergebnisse. Hervorzuheben ist die Teilnahme zweier Doktoranden am AAAI Spring Symposium in Kalifornien.

Ein weiterer Posten entfiel auf die Vergabe eines Forschungsauftrages (Kostenstelle 0835) zur Distanzschätzung mittels Stereo- und Bildinformationen an Herrn Lutz Schäfer. Die Ergebnisse wurden in Kooperation mit Siemens Mobility (SMO) und ITQ genutzt, um Modelle mit realistischeren Tiefenschätzungen zu trainieren.

Erzieltes Ergebnis

Durch den gezielten Einsatz der Fördermittel konnten die geplanten wissenschaftlichen Arbeiten im Projekt *safeTrAln* erfolgreich umgesetzt werden. Insbesondere wurde erreicht:

- Datenqualität und Absicherung: Entwicklung und Erprobung neuer Metriken zur Datenqualität (ECS, QI²) sowie darauf aufbauender Absicherungsalgorithmen. Ergänzend wurden ressourcenschonende Verfahren zur Datenqualitätsprüfung ohne vorheriges Training entwickelt und chaostheoretische Ansätze zur Stabilitätsprüfung untersucht. Die Methoden ermöglichten eine differenzierte Bewertung von Daten, die Identifikation von Edge Cases sowie die systematische Untersuchung der Robustheit von KI-gestützten Objekterkennungs- und Segmentierungsverfahren.
- Optische Modellierung und Sensorik: Entwicklung physikalisch-realistisch angepasster Optikmodelle für Kamera-Sensoren, um den Einfluss optischer Eigenschaften auf Bildqualität, Sensorfusion und KI-Modelle zu analysieren. Ergänzend erfolgte die Untersuchung optischer Unsicherheiten, insbesondere von Mess- und Kalibrierdrift, und deren Auswirkungen auf die Robustheit von KI-Methoden.
- Sensorische Optimierung: Mitarbeit an der Definition und Spezifikation der optischen Sensorik und damit an der konzeptionellen Ausgestaltung des Sensorsystems. Anstelle des ursprünglich geplanten Prüfstands konnten durch die Optimierung extrinsischer und intrinsischer Kameraparameter substantielle Verbesserungen erzielt und alternative Wege zur Steigerung der Modellgüte erschlossen werden.
- Wissenschaftliche Sichtbarkeit: Veröffentlichung und Präsentation der Ergebnisse auf internationalen Konferenzen und in Fachzeitschriften (u. a. *AI & Ethics*, AAAI Spring Symposium, Einreichungen bei CVPR und WACV). Die internationale Dissemination stärkte die Sichtbarkeit der Hochschule Düsseldorf und festigte ihre Rolle in der Forschungsgemeinschaft.
- Kooperationen und Transfer: Beiträge zur *Landscape of Safety Concerns* (Siemens Mobility) sowie zur Entwicklung von GSN-Bäumen (Fraunhofer IKS). Der Forschungsauftrag zur Distanzschätzung lieferte praxisnahe Ergebnisse für industrielle Anwendungen und verbesserte die Trainingsdaten für KI-Modelle durch realistischere Tiefenschätzungen.
- Nachwuchsförderung: Drei Promotionsstellen wurden eingerichtet, welche substantiell

zum Projektfortschritt beitragen. Zwei Promovierende setzen ihre Arbeiten in Kooperation mit der Goethe-Universität Frankfurt fort und stehen kurz vor dem Abschluss. Zusätzlich konnten drei wissenschaftliche Hilfskräfte eingebunden und insgesamt neun Abschlussarbeiten betreut werden.

Insgesamt wurden die Fördermittel zweckentsprechend eingesetzt und führten zu methodischen, technischen, wissenschaftlichen und personellen Ergebnissen von nachhaltigem Wert. Neben der wissenschaftlichen Verwertung wurde durch die Nachwuchsförderung, die Entwicklung neuer Methoden und die Zusammenarbeit mit Industriepartnern eine tragfähige Basis für künftige Forschungsaktivitäten geschaffen.

2. Wichtigste Positionen des Zahlenmäßigen Nachweises

Die wichtigsten Positionen sind vor allem die 0812, 0817 und 0822. Diese decken die Personalkosten ab.

Die Position 0812 war geplant für 3 vollfinanzierte Promotionsstellen. Allerdings wurden hiervon lediglich zwei volle Stellen über die volle Länge des Projektes 01/22 bis 03/25, eine volle Stelle über den Zeitraum 05/22 bis 12/24 und eine halbe Stelle über den Zeitraum 01/23 bis 06/23 genutzt.

Die Positionen 0817 und 0822 waren zur Unterstützung durch wissenschaftliche und studentische Hilfskräfte angedacht. Allerdings wurden auch hierbei nur eine halbe Stelle im Zeitraum 03/22 bis 08/23, eine viertel Stelle im Zeitraum 11/23 bis 03/25 und eine Stelle auf Basis von 10h im Zeitraum 07/24 bis 12/24 genutzt.

Die Position 0835 wurde für externe Auftragsarbeit im Bereich der Sensorik und Optik verwendet. Damit wurde eine Beratung im Zeitraum 07/24 bis 03/25 abgedeckt.

Die Position 0846 war für Dienstreisen vor allem im Inland plus eine Auslandsreise für einen Konferenzbeitrag pro Promotion gedacht. Inländische Dienstreisen wurden überwiegend für Teilnahmen an Konsortialtreffen genutzt. Zusätzlich wurden je eine außer-Europäische Reise für Konferenzbeiträge von zwei Promotionen genutzt (siehe Resiebericht).

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die im Projekt erbrachten Arbeiten waren in hohem Maße notwendig und angemessen, um die Projektziele im Hinblick auf die sichere Nutzung KI-basierter Systeme im Bahnumfeld zu erreichen. Eine Schlüsselrolle kam dabei der Absicherung der eingesetzten Verfahren zu. Insbesondere galt es, die Überwachung und Bewertung der Datenqualität, die Erklärbarkeit der KI-Komponenten sowie die Untersuchung der Sensitivität gegenüber Störeinflüssen sicherzustellen. Diese Aspekte bilden zentrale Bausteine einer belastbaren Sicherheitsargumentation und waren daher essenziell für die Projektdurchführung.

Darüber hinaus zeigte sich, dass die Auslegung der Optik und die Qualität der Kamerasysteme einen direkten Einfluss auf die Zuverlässigkeit der KI-Komponenten haben. Auch wenn nicht alle geplanten Arbeiten im Bereich der Kamera-Simulation umgesetzt werden konnten, wurden mit der dynamischen extrinsischen Kalibrierung, der Auslegung der Sensorik sowie den Testdatenerhebungen auf dem HVLE-Testgelände zentrale Arbeiten erbracht, die für die sichere Funktionsweise der Systeme unerlässlich sind. Diese optikbezogenen Untersuchungen lieferten eine wichtige Grundlage für die spätere algorithmische Weiterverarbeitung.

Ein weiterer Schwerpunkt lag auf der Integration der entwickelten Methoden in einen Gesamtkontext, der konsequent auf sicherheitsrelevante Kriterien ausgerichtet war. Hierbei konnten nicht nur bestehende Sicherheitsanforderungen adressiert, sondern auch eigene Impulse zur Mitgestaltung von Sicherheitskriterien gesetzt werden. Durch diese Arbeiten wurde eine solide Basis geschaffen und ihre Ergebnisse vertrauenswürdig nutzbar zu machen.

4. Voraussichtliches Nutzen im Sinne des Verwertungsplans

Wissenschaftliche

Verwertung

Das Projekt leistete einen wesentlichen Beitrag zur wissenschaftlichen Qualifizierung und Nachwuchsförderung. Geplant war die Ausbildung von drei Promotionsstellen, von denen zwei erfolgreich weitergeführt werden konnten. Beide Promovierende stehen kurz vor dem Abschluss und führen ihre Arbeiten in kooperativer Betreuung mit der Goethe-Universität Frankfurt durch. Dadurch ist nicht nur die inhaltliche Weiterentwicklung im Bereich der KI-Sicherheit gewährleistet, sondern auch eine institutionelle Verankerung an einer etablierten Universität. Eine dritte Promotionsstelle musste nach Abschluss des Projektes abgebrochen werden, was jedoch die Gesamtentwicklung nicht beeinträchtigte.

Ein zentraler wissenschaftlicher Nutzen liegt in der Entwicklung und Publikation von neuen Methoden und Metriken zur Bewertung von Datenqualität. Insbesondere die Metriken ECS und QI^2 wurden im Projekt konzipiert, angewandt und erprobt, sowie in der internationalen Fachwelt vorgestellt. Sie fanden Aufnahme in Beiträge beim AAAI Spring Symposium sowie in der Fachzeitschrift *AI & Ethics* (Springer) und wurden bereits in weiteren wissenschaftlichen Arbeiten zitiert und angewendet. Die Nutzung dieser Metriken ermöglicht eine deutlich differenziertere Bewertung von Trainings- und Testdaten in KI-Systemen und adressiert damit ein zentrales Problem im Bereich sicherheitskritischer Anwendungen.

Darüber hinaus wurden im Projekt neuartige Absicherungsalgorithmen entwickelt, die QI^2 mit etablierten Verfahren der Erklärbarkeit, wie Layer-wise Relevance Propagation (LRP), kombinieren. Diese Methoden wurden getestet, bewertet und in mehreren Publikationen auf internationalen Konferenzen eingereicht. Während ein Beitrag zur CVPR trotz positiver Gutachterkommentare zur Neuheit und methodischen Güte abgelehnt wurde, befindet sich eine erweiterte Fassung in Begutachtung bei der WACV. Parallel dazu werden die Verfahren auf verschiedene Netzarchitekturen übertragen, zur Automatisierung weiterentwickelt und in Hinblick auf die Qualität der Ausgabe kontinuierlich optimiert. Neben dem Verfahren zur Nutzung des QI^2 in der Erklärbarkeit von neuronalen Netzen, wurde auch der ECS genutzt, um Transparenz zu erschaffen. Diese Methode wurde ebenfalls bei der CVPR eingereicht und erhielt mit positivem Feedback ebenso eine Absage. Zusätzlich wurde ein Verfahren zur Datenqualitätsprüfung auf Basis von Feature-Ähnlichkeiten ohne Training entwickelt. Dieser Ansatz erlaubt eine schnelle und ressourcenschonende Identifikation fehlerhafter Datenpunkte in Trainings- und Testdatensätzen, ohne dass ein Modell zuvor aufwendig trainiert werden muss. Erste Veröffentlichungen hierzu wurden ebenfalls bei führenden Konferenzen eingereicht (WACV). Die Methode wurde erfolgreich auf State-of-the-Art-Datensätze angewandt und zeigt großes Potenzial für den praktischen Einsatz in der Datenkuratierung.

Ein zusätzlicher nächster Forschungsstrang beschäftigte sich mit einem chaostheoretischen Ansatz zur Stabilitätsprüfung von KI-Systemen, der auf den Analysen mit QI^2 und LRP aufbaut. Hieraus wird im weiteren Verlauf eine weitere Veröffentlichung generiert. Die Methodik um die Datenqualitätsprüfung wird parallel dazu ebenso weiterentwickelt, indem Komponenten auf hinreichende Nutzbarkeit überprüft und ausgetauscht, bzw. weiterentwickelt werden. Unter anderem werden auch weitere Nutzbarkeit dieser Methodiken in anderen Bereichen erforscht.

Ein weiterer Beitrag zur wissenschaftlichen Verwertung ist die Betreuung von insgesamt neun Abschlussarbeiten teilweise in Kooperation mit Siemens Mobility. Diese Arbeiten behandelten unterschiedliche Schwerpunkte: die Weiterentwicklung von Absicherungsmethoden für KI, die Entwicklung eines RBF-Klassifikators, die Implementierung physikalisch-realistischer Degradationen in Simulationen sowie die automatisierte Fahrwegsannotation. Diese Abschlussarbeiten ermöglichten es, Teilaspekte des Projekts vertiefend zu bearbeiten und zugleich Studierende frühzeitig an das Forschungsumfeld heranzuführen.

Wissenschaftliche

Anschlussfähigkeit

Die im Projekt erzielten Ergebnisse eröffnen eine breite Anschlussfähigkeit für weitere Forschungsarbeiten. Besonders hervorzuheben ist die Neubesetzung einer Promotionsstelle im Bereich Optik und Sensorik, die sich mit der Real-Time Optical Degradation und deren Auswirkung auf die Leistungsfähigkeit von KI-Systemen befasst. Damit wird die Brücke geschlagen zwischen physikalischen Eigenschaften der Sensorik und der Robustheit datengetriebener Modelle.

Darüber hinaus wurden im Projekt Methodenbaukästen zur Daten- und KI-Absicherung

entwickelt, die als Grundlage für neue Forschungsvorhaben dienen. Diese Methodiken sollen in weiteren Projekten angewendet, erweitert und in Kooperationen sowohl innerhalb der Hochschule Düsseldorf als auch mit externen Partnern (z. B. Industrie und Forschungsinstitute) verbreitet werden.

Zukünftige Forschungsrichtungen umfassen insbesondere:

- die Weiterentwicklung deterministischer Fahrwegserkennung mit Unterstützung durch KI, um robuste und erklärbare Trajektorienplanung zu ermöglichen,
- die Kombination von Absicherungsmethoden mit optischen Systemen, um die Sensitivität gegenüber realen sensorischen Einflüssen besser zu adressieren, sowie
- die Übertragung der entwickelten Absicherungsalgorithmen auf verschiedene KI-Architekturen und deren Integration in standardisierte Nachweisprozesse.
- Automatisierung der Absicherungsalgorithmen
- Weiterentwicklung der Methoden zur Absicherung von KI-Systemen und der Datenqualität

Damit trägt das Projekt nicht nur zur wissenschaftlichen Diskussion über KI-Sicherheit bei, sondern liefert auch konkrete methodische und personelle Grundlagen für zukünftige Forschung. Es schafft die Basis für eine nachhaltige Verwertung sowohl in Form von wissenschaftlichen Publikationen und Qualifikationsarbeiten als auch durch die Fortführung in weiterführenden Projekten und Kooperationen.

5. Fortschritt bei anderen Stellen während des Vorhabens

Die vergangenen Jahre waren von erheblichen technologischen Entwicklungen im Bereich der Künstlichen Intelligenz geprägt. Besonders sichtbar wurde dies im Kontext von Sprachmodellen (LLMs), deren Leistungsfähigkeit und Verbreitung stark zugenommen hat. Diese Fortschritte betreffen jedoch in erster Linie allgemeine Anwendungen und konzentrierten sich nicht auf die Absicherung von KI-Systemen oder den Bahnbetrieb.

Neben diesen allgemeinen Entwicklungen konnten bei einzelnen externen Forschungsstellen eine Reihe spezialisierter, forschungstechnisch relevanter Neuerungen beobachtet werden. Ein Beispiel hierfür sind Arbeiten zur Verbesserung der Datenqualitätsuntersuchung, die sich in den letzten Jahren von überwiegend klassifikationsbasierten Ansätzen hin zu detektionsorientierten Verfahren weiterentwickelt haben [16], [17], [18]. Vergleichbare Fortschritte wurden auch in angrenzenden Bereichen erzielt, etwa bei der robusten Merkmalsextraktion, der Analyse von Unsicherheiten oder der Sensorfusion. Weitere Entwicklungen konnten auch im Themenbereich der Erklärbarkeit und Transparenz von KI-Systemen beobachtet werden. So wurden unter anderem post-hoc Methoden für Transformerbasierte Netzwerke (weiter-)entwickelt [19], [20], [21]. Aber auch Methoden für klassische Neuronale Netze wurden weiterentwickelt und auf aktuellere Anforderungen spezialisiert [22], [23], [24], [25]. Solche Entwicklungen wurden kontinuierlich verfolgt, evaluiert und – sofern zweckmäßig – in die eigene Arbeit integriert.

Insgesamt zeigt sich, dass die externe Forschung wertvolle Impulse geliefert hat, die den Projektfortschritt punktuell bereichert und die eigenen Ansätze gestützt haben. Eine grundlegende Neuausrichtung des Vorhabens war auf dieser Basis jedoch nicht erforderlich, vielmehr konnten die Arbeiten im Projekt durch die externe Entwicklung sinnvoll ergänzt und validiert werden.

6. Erfolgte und geplant Veröffentlichung der Ergebnisse

Im Rahmen des Projektes konnten bereits mehrere Veröffentlichungen erfolgreich realisiert werden. Dazu zählen insbesondere Beiträge mit Bezug zu den Queens-Methoden QI^2 und ECS, die sowohl als eigenständige Arbeiten als auch in weiterführenden Anwendungskontexten publiziert wurden. Beispiele hierfür sind die Veröffentlichungen *ECS – an Interactive Tool for Data Quality Assurance* [14] und *QI^2 – an Interactive Tool for Data Quality Assurance* [13], die bei Springer und der AAI erschienen sind. Ergänzend dazu wurden die entwickelten Methoden in die Arbeit *Continuous Development and Safety Assurance Pipeline for ML-Based Systems in the Railway Domain* [15] integriert und dort in einem breiteren Anwendungskontext vorgestellt.

Darüber hinaus befinden sich derzeit zwei weiterentwickelte methodische Ansätze in der wissenschaftlichen Begutachtung. Diese Arbeiten knüpfen an die bisherigen Ergebnisse an und erweitern diese um neue Verfahren zur Absicherung von KI-Modellen und Qualitätsbewertung von Datensätzen. Bei diesen Arbeiten handelt es sich um *CRISP – Complexity-based Reasoning of Internal Subprocessing* sowie *Training-Free Identification of Annotation Corruptions in Object Detection Datasets*.

Insgesamt konnten im Rahmen des Projektes sowohl bereits sichtbare Ergebnisse hervorgebracht als auch noch laufende Beiträge erzeugt werden. Damit ist die Grundlage gelegt, die erarbeiteten Methoden über die Projektdauer hinaus in der wissenschaftlichen Community zu verankern und weiterzuentwickeln.

Teil III:

1. Beitrag des Ergebnisses zu den förderpolitischen Zielen

Das Projekt safetrAln ist im Rahmen der Programmsäule „Automatisiertes Fahren“ des Förderprogramms für den Schienenverkehr angesiedelt. Das Ziel dieser Programmsäule besteht in der Schaffung der Grundlagen für einen vollautomatisierten Zugbetrieb, insbesondere durch die Absicherung von KI-basierten Perzeptionssystemen. Dabei steht die sichere Erkennung von Hindernissen als zentrale technologische Herausforderung im Fokus, um den fahrerlosen Betrieb von Schienenfahrzeugen zuverlässig und zulassungskonform zu gestalten.

Im Rahmen des Projekts übernahm die HSD zentrale Aufgaben in der KI-Absicherung. Dabei lag der Schwerpunkt auf der Entwicklung und Weiterentwicklung der sogenannten Queens-Methoden, insbesondere QI^2 (integrierter Qualitätsindikator) und ECS (equivalent class sets). Diese Methoden ermöglichten eine systematische Analyse der Datenqualität innerhalb der Trainings- und Testdatensätze und erlaubten zugleich eine fundierte Untersuchung der Robustheit komplexer KI-Modelle in der Objekterkennung und Segmentierung. Darüber hinaus konnten mit diesen Methoden gezielt Edge Cases identifiziert und Unsicherheiten in der KI-gestützten Objekterkennung propagiert und nachverfolgt werden.

Die entwickelten Queens-Methoden fanden darüber hinaus Anwendung bei der Validierung von RBF-Netzwerken und der Generierung geprüfter Datensätze, wodurch sie einen direkten Beitrag zur Absicherung von KI-Systemen leisteten.

Im Bereich der optischen Modellierung und Sensorik konzentrierte sich die HSD auf die Entwicklung physikalisch realistischer Modelle für Kamera-Sensoren. Ziel war es, die Einflüsse optischer Eigenschaften auf die Bildqualität, auf die Sensorfusion und auf die Leistungsfähigkeit der KI-gestützten Objekterkennung systematisch zu untersuchen. Dabei wurden insbesondere Unsicherheiten wie Mess- und Kalibrierdrift analysiert, um ihre Auswirkungen auf die Robustheit der KI-Methoden abzuschätzen. Ergänzend trug die HSD zur Definition und Spezifikation der optischen Sensorik bei und lieferte damit wichtige Impulse für die konzeptionelle Gestaltung des Sensorsystems. Obwohl der ursprünglich geplante Prüfstand nicht zum Einsatz kam, konnten die extrinsischen und intrinsischen Kameraparameter optimiert werden, wodurch eine alternative Möglichkeit zur Verbesserung der Modellgüte gebildet wurde.

Weitere Arbeiten der HSD fokussierten sich auf die Unterstützung bei der Absicherung der KI-Modelle. In Zusammenarbeit mit Fraunhofer IKS arbeitete die HSD an der Entwicklung von GSN-Bäumen (Goal Structuring Notation) und in Zusammenarbeit mit SMO an der ‚Landscape of safety concerns‘ in den relevanten Themengebieten. In diesen Arbeiten wurde sicherheitskritisch Unterpunkte ausgearbeitet, mit zugehörigen Methoden und Metriken versehen und, sofern möglich, validiert.

Im Bereich der wissenschaftlichen Verwertung engagierte sich die HSD, um die erzielten Projektergebnisse nachhaltig in die Forschungsgemeinschaft einzubringen. Dazu gehörte die Betreuung kooperativer Promotionen sowie die Veröffentlichung und Präsentation der Ergebnisse in mehreren wissenschaftlichen Publikationen (siehe „Teil II, 6“). Es ist geplant die entwickelten Methoden und Modelle auch in weiteren Forschungsvorhaben zu adaptieren und weiterzuentwickeln.

Die Arbeiten der HSD trugen zur Absicherung der KI-Modelle bei, unter anderem durch die Anwendung der Queens-Methoden sowie der entwickelten optischen Modelle. Diese Methoden ermöglichten eine systematische Untersuchung der Datenqualität und der Robustheit der eingesetzten Verfahren sowie eine Bewertung des Einflusses optischer Eigenschaften auf die Sicherheitsarchitektur. Darüber hinaus wurden die Projektergebnisse in der wissenschaftlichen Gemeinschaft eingebracht, unter anderem durch Publikationen und die Betreuung kooperativer Promotionen, und bieten eine Grundlage für weitere Forschungsaktivitäten.

2. Wissenschaftlich-technisches Ergebnis des Vorhabens

Im Projekt wurden verschiedene Methoden entwickelt, um die Qualität von Datensätzen zu bewerten und die interne Verarbeitung von KI-Modellen nachvollziehbar zu gestalten. Hierzu zählen insbesondere die Queens-Methoden (QI² und ECS). Der ECS (Equivalent Class Set), basiert auf der Analyse von Datenpunkt-Nachbarschaften, während die Methode QI², eine systematische Bewertung der Datenqualität und der Komplexität von Datensätzen ermöglicht. Eine weitere Erweiterung, die QI² Network Analysis, überträgt diesen Ansatz auf die Analyse neuronaler Netze, um die Verarbeitungsstrategie innerhalb der Modelle zu untersuchen. Diese Methoden dienen als Grundlage für die Absicherung und Weiterentwicklung der eingesetzten KI-Systeme und wurden in Fact-Sheets detailliert dargestellt.

Die HSD unterstützte bei der Auswahl geeigneter Kamerasysteme und das Testmodell der HVLE. In diesem Rahmen wurden auch Messungen auf dem Prüfgelände der HVLE durchgeführt. Des Weiteren wurde eine Methode für die dynamische Extrinsische Kalibrierung der Kamerasysteme entwickelt, welche auf der bekannten Position der Kamera sowie der größtenteils bekannten Position der Schiene im Bild basierte. Die hier erzielten Ergebnisse konnten bislang noch nicht in wissenschaftlichen Publikationen veröffentlicht werden.

Ein weiterer Beitrag des Projekts lag in der Mitarbeit an der Sicherheitsarchitektur. Hierbei unterstützte die HSD die Nachweisführung sicherheitsrelevanter Aspekte und trug zur Bewertung der Einflüsse optischer Eigenschaften auf RAMS- sowie nicht-funktionale Anforderungen einerseits und der Erstellung der GSN-Bäume sowie mithilfe bei der ‚Landscape of safety Concerns‘ andererseits bei. Darüber hinaus wirkte die HSD an der Erarbeitung des DIN DKE SPEC 99002 „Terminology: AI in Railway Applications“ mit und leistete damit einen Beitrag zur Standardisierung zentraler Begriffe im Themenfeld der Künstlichen Intelligenz im Bahnumfeld.

Darüber hinaus brachte die HSD methodische Erweiterungen in das Projekt ein. Neben der Unterstützung bei der Arbeit mit RBF-Netzwerken wurden insbesondere die im Projekt entwickelten Queens-Methoden dem Konsortium bereitgestellt. Diese Methoden erweiterten die Grundlage für die gemeinsame Analyse von Datenqualität und Modellverhalten und wurden in verschiedenen Arbeitspaketen angewendet.

Die Ergebnisse des Projekts wurden von der HSD durch Publikationen und Präsentationen in die wissenschaftliche Gemeinschaft eingebracht und durch die Betreuung kooperativer Promotionen weiter vertieft. Darüber hinaus besteht der Wunsch und die Möglichkeit, die entwickelten Ansätze in künftigen Forschungsvorhaben weiterzuführen, zu entwickeln und zu verfeinern.

3. Fortschreibung des Verwertungsplans

Wissenschaftliche Erfolgsaussichten und Anschlussfähigkeit

- Forschung und Entwicklung der Methoden dient der wissenschaftlichen Ausbildung der Promotionsstellen
 - o Neuer Beitrag zum aktuellen Stand der Wissenschaft um KI und Daten
 - o Bereitstellung der Ergebnisse für die Forschung
 - o Erweiterung auf Transformerbasierte Netzwerke
 - o Anwendung der Methoden auf Datenbanken aus der Forschung
- Mögliche Zusammenarbeit mit Fraunhofer HHI → Weiterentwicklung zwischen QI² und LRP (über das nächste Jahr ~ Mitte 2026)
- Mögliche Zusammenarbeit mit Arbeitsgruppe des KIT → chaostheoretische Untersuchung von KI (über das nächste Jahr ~ Ende 2026)
- Weiterentwicklung der erstellten Methoden und veröffentlichen im Vergleich zum Stand der aktuellen Forschung

Die im Projekt erzielten Ergebnisse wurden in der wissenschaftlichen Ausbildung von Doktoranden sowie Wissenschaftlichen Mitarbeitern im Master- und Bachelorstudium genutzt. Auf diese Weise konnte das Vorhaben direkt in die Nachwuchsförderung eingebunden werden. Zudem wurden die entwickelten Methoden für wissenschaftliche Publikationen aufbereitet, um die Forschungsgemeinschaft an den erzielten Fortschritten teilhaben zu lassen.

Hinsichtlich der wissenschaftlichen Erfolgsaussichten haben die Arbeiten neue Beiträge zum

aktuellen Stand der Forschung im Bereich Künstliche Intelligenz und Datenanalyse hervorgebracht. Die entwickelten Ansätze bieten eine Grundlage für Anwendungen auf weitere Forschungsdatenbanken und lassen sich perspektivisch auch auf transformerbasierte Netzwerke übertragen. Damit eröffnen sich Chancen für eine kontinuierliche Weiterentwicklung im Vergleich zum Stand der Forschung.

Über das Projekt hinaus bestehen Anknüpfungspunkte für weitere Kooperationen. Dazu zählt eine mögliche Zusammenarbeit mit dem Fraunhofer HHI zur Verknüpfung von QI² und LRP sowie ein Austausch mit einer Arbeitsgruppe des KIT zur chaostheoretischen Untersuchung von KI-Systemen. Diese Vorhaben sollen im Zeitraum bis 2026 verfolgt werden und tragen dazu bei, die Anschlussfähigkeit der erzielten Ergebnisse auch nach Abschluss des Projekts zu sichern.

4. Arbeiten, die zu keiner Lösung geführt haben

Grundsätzlich konnte im Projekt für alle bearbeiteten Schwerpunkte eine Lösung erarbeitet werden. Lediglich einige ursprünglich angedachte Arbeiten sind im Verlauf zurückgestellt worden. Dazu zählt insbesondere die Idee eines Runtime Monitorings auf Basis von QI², die aus Ressourcengründen nicht weiter umgesetzt wurde.

Auch im Bereich Optik und Sensorik konnten einige Ansätze nicht zur geplanten Veröffentlichung geführt werden, obwohl konzeptionelle Arbeiten vorlagen. Hierzu zählen insbesondere die dynamische extrinsische Kalibrierung sowie Ansätze zur deterministischen, KI-unterstützten Fahrwegserkennung. Diese Themen behalten ihre wissenschaftliche Relevanz und stellen potenzielle Anschlussarbeiten dar, wurden im Rahmen des Vorhabens jedoch nicht vollständig umgesetzt.

5. Präsentationsmöglichkeiten für mögliche Nutzer

Die im Rahmen des Vorhabens entwickelten Methoden und Ergebnisse der HSD eignen sich in erster Linie für die Präsentation auf wissenschaftlichen Konferenzen im Bereich Künstliche Intelligenz, Datenqualität und Erklärbarkeit von Modellen. Hier können sie einem internationalen Fachpublikum vorgestellt und im Kontext des aktuellen Forschungsstands diskutiert werden.

Darüber hinaus besteht die Möglichkeit, die Arbeiten in domänenspezifischen Veranstaltungsformaten mit Bezug zur Bahn zu präsentieren, sofern ein thematischer Schwerpunkt auf die Anwendung von KI im Schienenverkehr gelegt wird. Dies betrifft insbesondere Konferenzen und Workshops, die sich mit Fragen der Digitalisierung, Sicherheit und Datenverarbeitung im Bahnumfeld befassen.

6. Einhaltung der Kosten- und Zeitplanung

Es wurden der Kosten-, sowie der vorliegende Zeitplan eingehalten. Zusätzlich wurde, wie für das gesamte Konsortium eine Verlängerung um 3 Monate beantragt und inhaltlich sinnvoll durchgeführt. Diese Verlängerung galt der abschließenden Implementierung, Anwendung und dem Sammeln von Ergebnissen, die einem hinreichenden Projektabschluss dienen.

Referenzen

- [1] European Commission, LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. 2021.
- [2] O. Willers, S. Sudholt, S. Raafatnia, and S. Abrecht, "Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks," Jan. 22, 2020, arXiv: arXiv:2001.08001. Accessed: Dec. 27, 2022. [Online]. Available: <http://arxiv.org/abs/2001.08001>
- [3] T. Waschulzik, "Qualitätsgesicherte effiziente Entwicklung vorwärtsgerichteter künstlicher Neuroner Netze mit überwachtem Lernen (QUEEN)," Technische Universität München, München, 1999.
- [4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PLoS ONE, vol. 10, no. 7, p. e0130140, Jul. 2015, doi: 10.1371/journal.pone.0130140.
- [5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," Apr. 19, 2014, arXiv: arXiv:1312.6034. Accessed: Aug. 09, 2023. [Online]. Available: <http://arxiv.org/abs/1312.6034>
- [6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.
- [7] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV: IEEE, Mar. 2018, pp. 839–847. doi: 10.1109/WACV.2018.00097.
- [8] P. Müller and A. Braun, "MTF as a performance indicator for AI algorithms?," in Electronic Imaging, Jan. 2023, pp. 125-1-125–1. doi: 10.2352/EI.2023.35.16.AVM-125.
- [9] P. Müller and A. Braun, "Local performance evaluation of AI-algorithms with the generalized spatial recall index," tm - Technisches Messen, vol. 0, no. 0, May 2023, doi: 10.1515/teme-2023-0013.
- [10] P. Müller, A. Braun, and M. Keuper, "Impact of realistic properties of the point spread function on classification tasks to reveal a possible distribution shift," in NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications, Dec. 2022. [Online]. Available: <https://openreview.net/forum?id=r7WJpE3oy0>
- [11] P. Müller, A. Braun, and M. Keuper, "Examining the Impact of Optical Aberrations to Image Classification and Object Detection Models," Apr. 25, 2025, arXiv: arXiv:2504.18510. doi: 10.48550/arXiv.2504.18510.
- [12] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks," Nov. 07, 2021, arXiv: arXiv:2103.14749. doi: 10.48550/arXiv.2103.14749.
- [13] S. Geerkens, C. Sieberichs, A. Braun, and T. Waschulzik, "QI²: an interactive tool for data quality assurance," AI and Ethics, Jan. 2024, doi: 10.1007/s43681-023-00390-6.
- [14] C. Sieberichs, S. Geerkens, A. Braun, and T. Waschulzik, "ECS: an interactive tool for data quality assurance," AI and Ethics, Jan. 2024, doi: 10.1007/s43681-023-00393-3.
- [15] M. Zeller et al., "Continuous Development and Safety Assurance Pipeline for ML-Based Systems in the Railway Domain," in Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops, A. Ceccarelli, M. Trapp, A. Bondavalli, E. Schoitsch, B. Gallina, and F. Bitsch, Eds., Cham: Springer Nature Switzerland, 2024, pp. 446–459.
- [16] U. Tkachenko, A. Thyagarajan, and J. Mueller, "ObjectLab: Automated Diagnosis of Mislabeled Images in Object Detection Data." 2023. [Online]. Available: <https://arxiv.org/abs/2309.00832>
- [17] K. Chachula, J. Łyskawa, B. Olber, P. Frątczak, A. Popowicz, and K. Radlak, "Combating noisy labels in object detection datasets." 2023. [Online]. Available: <https://arxiv.org/abs/2211.13993>
- [18] Z. Zhu, Z. Dong, and Y. Liu, "Detecting Corrupted Labels Without Training a Model to Predict." 2022. [Online]. Available: <https://arxiv.org/abs/2110.06283>
- [19] R. Achibat et al., "AttnLRP: Attention-Aware Layer-Wise Relevance Propagation for

- Transformers,” in ICML, 2024. [Online]. Available: <https://openreview.net/forum?id=emtXYIBrNF>
- [20] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso, “Towards Automated Circuit Discovery for Mechanistic Interpretability,” Oct. 07, 2023, arXiv: arXiv:2304.14997. Accessed: Oct. 24, 2023. [Online]. Available: <http://arxiv.org/abs/2304.14997>
- [21] C. Zhao, J. H. Hsiao, and A. B. Chan, “Gradient-Based Instance-Specific Visual Explanations for Object Specification and Object Discrimination,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 9, pp. 5967–5985, 2024, doi: 10.1109/TPAMI.2024.3380604.
- [22] Y. Wu, C. Chen, J. Che, and S. Pu, “FAM: Visual Explanations for the Feature Representations From Deep Convolutional Networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA: IEEE, Jun. 2022, pp. 10307–10316. [Online]. Available: <https://ieeexplore.ieee.org/document/9879356/>
- [23] R. Achibat et al., “From attribution maps to human-understandable explanations through Concept Relevance Propagation,” Nat Mach Intell, vol. 5, no. 9, pp. 1006–1019, Sep. 2023, doi: 10.1038/s42256-023-00711-8.
- [24] N. Akhtar and M. A. A. K. Jalwana, “Rethinking Interpretation: Input-Agnostic Saliency Mapping of Deep Visual Classifiers,” AAAI, vol. 37, no. 1, pp. 178–186, Jun. 2023, doi: 10.1609/aaai.v37i1.25089.
- [25] J. Vielhaben, S. Blücher, and N. Strodthoff, “Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees,” Jun. 18, 2023, arXiv: arXiv:2301.11911. Accessed: Oct. 10, 2023. [Online]. Available: <http://arxiv.org/abs/2301.11911>