safe.trAIn
Sichere KI am Beispiel fahrerloser Regionalzug

Abschlussbericht

Zuwendungsempfänger
Otto von Guericke Univerisität Magdeburg
Universitästplatz 2, 39106 Magdeburg

Projektleitung:
Prof. Dr. rer. nat. Frank Ortmeier

Förderkennzeichen: 19I21039N
Projektlaufzeit: 01.01.2022-31.03.2025

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 19I21039N gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Table of Contents

I Abschlussbericht - Individueller Schlussbericht

1 Aufgabenstellung

In den vergangenen Jahren haben Methoden der Künstlichen Intelligenz (KI), insbesondere im Bereich des maschinellen Lernens (ML), bemerkenswerte Fortschritte erzielt. Diese Entwicklung betrifft zunehmend auch sicherheitskritische Systeme wie beispielsweise autonome Fahrzeuge, medizinische Diagnoseinstrumente und automatisierte Schienenfahrzeuge. Gleichzeitig steigt das gesellschaftliche Interesse an der verstärkten Automatisierung von Verkehrsmitteln, um Emissionen zu reduzieren und dem Klimawandel entgegenzuwirken. Besonders im Schienenverkehr ist eine Ausweitung des vollautomatisierten Betriebs (GoA4) erforderlich, um attraktive, flexible und kostengünstige Mobilitätsangebote bereitstellen zu können.

Allerdings gibt es bislang erhebliche Herausforderungen bei der Zertifizierung und Sicherheitsnachweisführung KI-basierter Systeme. Die Anwendung dieser Technologien im sicherheitskritischen Bereich wirft Fragen bezüglich ihrer Zuverlässigkeit, Erklärbarkeit, Robustheit und regulatorischen Zulassung auf. Insbesondere fehlen standardisierte Prüfmethoden und Bewertungsverfahren für KI-Funktionen, was ihre Integration in bestehende Zulassungsprozesse erschwert. Das safe.trAIn-Vorhaben adressierte genau diese Herausforderungen. Ziel des Projekts war es, methodische und technologische Grundlagen für den sicheren und regulatorisch akzeptierten Einsatz von KI-Methoden am Beispiel eines fahrerlosen Regionalzugs zu schaffen. Ein Schwerpunkt lag hierbei auf der Erforschung von Prüfmethoden und Werkzeugen zur Sicherheitsbewertung von KI, der Entwicklung einer geeigneten Sicherheitsarchitektur sowie der prototypischen Umsetzung und Evaluation in einem virtuellen Testfeld.

2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Vorhaben wurde im Rahmen des Fachprogramms „Neue Fahrzeug- und Systemtechnologien“ des Bundesministeriums für Wirtschaft und Energie durchgeführt und hatte eine Laufzeit von 36 Monaten (01.01.2022 – 31.12.2024, b.z.w. 39 Monate inklusive einer 3-Monatigen Kostenneutralen Verlängerung). Es bestand ein breites Konsortium aus Industriepartnern, Forschungseinrichtungen sowie Normungs- und Prüforganisationen. Die OVGU arbeitete eng mit Partnern wie Siemens, TÜV Rheinland, Fraunhofer IKS und weiteren Forschungseinrichtungen zusammen, um die Anforderungen und Spezifikationen gemeinsam abzustimmen und umzusetzen. Die Forschungsarbeiten wurden kontinuierlich an den aktuellen Stand der Technik sowie laufende Normungs- und Standardisierungsaktivitäten angelehnt, um die Anschlussfähigkeit der Projektergebnisse sicherzustellen.

3 Planung und Ablauf des Vorhabens

Das Vorhaben gliederte sich in mehrere Arbeitspakete (APs), in denen die OVGU spezifische Aufgaben übernahm.

3.1 AP 1: Anforderungen an die Sicherheitsnachweisführung

Die zentrale Herausforderung beim Einsatz KI-basierter Systeme im sicherheitskritischen Schienenverkehr liegt in der Entwicklung und Etablierung geeigneter Sicherheitsnachweisverfahren. Anders als bei klassischen deterministischen Systemen zeichnet sich zeitgenössische Künstliche Intelligenz, insbesondere das maschinelle Lernen, durch nicht-deterministische Eigenschaften, probabilistische Modellbildung und datengetriebene Entwicklungsmethoden aus. Diese Charakteristika erschweren die Anwendung etablierter sicherheitstechnischer Normen, wie etwa der EN 50129 oder der IEC 61508, erheblich.

Vor diesem Hintergrund übernahm die Otto-von-Guericke-Universität Magdeburg (OVGU) eine unterstützende Rolle bei der Erarbeitung und Analyse bestehender Normen, Standards und regulatorischer Anforderungen. Gemeinsam mit den Projektpartnern, insbesondere dem TÜV Rheinland, Siemens und Fraunhofer IKS, wurden einschlägige Regelwerke aus verwandten Industriebereichen – beispielsweise der Automobilindustrie (ISO 26262), der Luftfahrt (ARP4754A, DO-178C) sowie der Medizintechnik (ISO 13485) – identifiziert, auf ihre Übertragbarkeit hin analysiert und in strukturierter Form dokumentiert.

Ein besonderer Fokus lag dabei auf der Frage, inwiefern bestehende Zertifizierungs- und Nachweisverfahren für Softwarekomponenten auf KI-basierte Komponenten angewendet oder erweitert werden können. Die OVGU unterstützte aktiv die Formulierung von Anforderungen an Entwicklungsprozesse, Nachweisketten und Prüfmethoden, welche spezifisch auf die Besonderheiten von KI-Systemen zugeschnitten sind. Dabei wurde insbesondere die Integration von Unsicherheitsmetriken, Transparenzanforderungen und Nachvollziehbarkeitskriterien in die Sicherheitsargumentation diskutiert und methodisch untermauert.

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

Im Mittelpunkt der Arbeiten der OVGU im Projekt safe.trAIn stand das Arbeitspaket 2, welches sich der Entwicklung und Validierung sicherer KI-Methoden widmete. Die OVGU leitete hierbei zwei zentrale Unterarbeitspakete: zum einen den Themenkomplex „Verifikation Künstlicher Intelligenz“, zum anderen „Beschreibung der Betriebsumgebung und Sicherstellung der Datenqualität“.

Zu Beginn des Projekts wurden umfangreiche Recherchen zum Stand der Technik durchgeführt. Diese zielten darauf ab, etablierte und neue Verfahren zur Validierung und Absicherung KI-basierter Modelle zu identifizieren, zu bewerten und ihre Anwendbarkeit im Kontext sicherheitskritischer Anwendungen im Schienenverkehr zu untersuchen. Im Rahmen dieser Arbeiten entwickelte die OVGU eine Sammlung quantitativer Metriken, mit denen sich die Robustheit, Erklärbarkeit und Unsicherheit von neuronalen Netzen objektiv bewerten lassen.

Ein Schwerpunkt lag auf der Entwicklung geeigneter Verfahren zur Out-of-Distribution Detection (OOD), also der Erkennung von Eingaben, die deutlich außerhalb des Trainingsverteilungsbereichs liegen. Diese Fähigkeit ist von zentraler Bedeutung für den Einsatz in sicherheitskritischen Systemen, da solche Eingaben zu unvorhersehbarem Verhalten der KI führen können. Die OVGU untersuchte und kombinierte verschiedene methodische Ansätze – darunter regelbasierte Verfahren, neuronale Unsicherheitsquantifizierung (z. B. durch Bayesianische Netze und Monte-Carlo-Dropout) sowie generative Modelle wie GANs oder Diffusionsmodelle – mit dem Ziel, robuste und transparente OOD-Erkennungsmechanismen zu etablieren.

Diese Methoden wurden sowohl auf synthetischen als auch realitätsnahen Simulationsdaten erprobt und iterativ weiterentwickelt. Ein besonderes Augenmerk galt dabei der Erklärbarkeit und Nachvollziehbarkeit der Vorhersagen. Ziel war es, nicht nur die Erkennungsleistung zu maximieren, sondern auch ein besseres Verständnis dafür zu schaffen, unter welchen Bedingungen das KI-System zuverlässig funktioniert – und wann es eben nicht tut.

Parallel dazu wurden die entwickelten Verfahren in das sogenannte Minimum Viable Product (MVP) integriert, das als Referenzimplementierung für den Sicherheitsnachweis diente. Die Metriken der OVGU wurden im Sicherheitsargumentationsbaum (GSN-Tree) verortet, sodass ihr Beitrag zur Gesamtsicherheitsbewertung transparent nachvollzogen werden kann.

Ein weiterer Beitrag der OVGU bestand in der Etablierung und Betreuung einer Arbeitspacketübergreifenden „Special Interest Group“ (SIG), die sich mit der formalen Beschreibung der Betriebsumgebung („Operational Design Domain“, ODD) befasste. Die korrekte Beschreibung der ODD ist eine wesentliche Voraussetzung für den Sicherheitsnachweis, da sie den Gültigkeitsbereich der entwickelten KI-Funktionalität definiert. Die OVGU leitete die Arbeiten zur Erstellung eines konsistenten, modularen Modells der Betriebsumgebung, welches sowohl in der Simulation als auch für reale Szenarien nutzbar ist. Dieses Modell wurde sukzessive verfeinert, an neue Anforderungen angepasst und in das Gesamtsystem integriert.

Die Ergebnisse dieser Arbeiten wurden auf mehreren Fachkonferenzen und Workshops veröffentlicht und vorgestellt – unter anderem auf der SafeComp-Konferenz, dem Workshop on AI Safety Engineering (WAISE) sowie der WACV. Eine dieser Arbeiten wurde

mit einem Best-Paper-Award ausgezeichnet, was die wissenschaftliche Relevanz der entwickelten Methoden unterstreicht.

3.3 AP 3: Fahrzeugarchitektur im GoA4-Betrieb mit Fokus auf sichere KI-basierte Funktionen

Im Arbeitspaket 3 unterstützte die OVGU die Entwicklung einer funktionalen, sicherheitsgerichteten Systemarchitektur für den autonomen Regionalzug. Ziel war es, die in AP 2 entwickelten KI-Methoden – insbesondere zur Objekterkennung und OOD-Erkennung – systematisch in die Architektur des Gesamtsystems einzubinden. Die OVGU brachte hier ihre Expertise in der strukturellen und funktionalen Sicherheitsmodellierung ein und arbeitete eng mit anderen Projektpartnern an der Definition sicherheitskritischer Funktionen, ihrer Zuordnung zu Systemkomponenten und der Ableitung entsprechender Safety-Anforderungen.

Ein besonderer Beitrag der OVGU bestand in der konzeptionellen Einbettung von Unsicherheitsmetriken und Erklärbarkeitsmodulen in das Perzeptionssystem. Die entwickelten Verfahren sollten nicht als isolierte Komponenten betrachtet werden, sondern integraler Bestandteil einer redundanten und fehlertoleranten Systemarchitektur sein. Dazu wurden alternative Systempfade und Verifikationsmechanismen konzipiert, welche im Fall unzureichender Vorhersagesicherheit automatisch aktiviert werden können.

Die OVGU koordinierte ihre architekturbezogenen Arbeiten eng mit den Projektpartnern, um sicherzustellen, dass alle entwickelten Komponenten und Sicherheitsmechanismen technisch umsetzbar und mit den weiteren Teilsystemen kompatibel waren. Durch die enge Abstimmung konnte ein kohärenter Architekturentwurf realisiert werden.

3.4 AP 4: Virtuelles Testfeld, Sicherheitsbewertung

Im vierten Arbeitspaket war die OVGU wesentlich an der Umsetzung der zuvor entwickelten Konzepte in ein simulationsbasiertes Testfeld beteiligt. Ziel war es, eine praxisnahe Umgebung zu schaffen, in der die entwickelten KI-Systeme unter kontrollierten Bedingungen validiert werden können. Die OVGU integrierte hierzu ihre Metriken zur Unsicherheitsbewertung, Segmentierungsqualität und OOD-Erkennung in die Testinfrastruktur und validierte diese im Zusammenspiel mit den Komponenten der Projektpartner.

Eine zentrale Herausforderung bestand darin, die entwickelten Verfahren an die unterschiedlichen Datenformate und Schnittstellen der Simulationsumgebung anzupassen. Hierzu war eine enge Abstimmung mit Siemens und weiteren Partnern notwendig. Die OVGU entwickelte eine Softwarebibliothek, die es erlaubt, die digitale Beschreibung der Betriebsumgebung automatisiert auszulesen und mit den realisierten Metriken zu verknüpfen. Diese Bibliothek diente als Grundlage für zahlreiche Analysen zur ODD-Abdeckung, Segmentierungsperformanz und Fehleridentifikation.

Die Implementierungen wurden in mehreren Iterationen verbessert und den Partnern zur Verfügung gestellt. Rückmeldungen flossen kontinuierlich in die Weiterentwicklung ein.

3.5 AP 5: Standardisierung und Verbreitung

Die Ergebnisse der OVGU flossen in eine Vielzahl von Disseminations- und Verwertungsaktivitäten ein. Neben mehreren wissenschaftlichen Publikationen in Fachzeitschriften und Konferenzen beteiligte sich die OVGU an der Erstellung von Standardisierungsdokumenten, darunter insbesondere die DIN SPEC 99002 („Terminology – AI in Railway Applications“) und DIN SPEC 99004 („MLOps für sicherheitskritische KI“). Die Universität brachte ihre wissenschaftlichen Erkenntnisse in die Normungsprozesse ein, um die regulatorische Anschlussfähigkeit der entwickelten Methoden zu sichern.

Darüber hinaus wurden zahlreiche studentische Arbeiten im Kontext des Projektes betreut. Die safe.trAIn-Ergebnisse flossen direkt in die Lehre ein und wurden in Seminaren, Projektgruppen und Abschlussarbeiten weiterentwickelt. Diese Arbeiten trugen dazu bei, Nachwuchswissenschaftlerinnen und -wissenschaftler an das Themenfeld „Sichere KI“ heranzuführen und das Wissen nachhaltig in der akademischen Ausbildung zu verankern.

Zusätzlich wurden Gespräche mit Industriepartnern über eine mögliche Weiterführung und Transferprojekte geführt. Besonders im Bereich der generativen KI und Robotik bestehen konkrete Anknüpfungspunkte für zukünftige Forschungs- und Entwicklungskooperationen.

4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

4.1 Angabe bekannter Konstruktionen, Verfahren und Schutzrechte, die für die Durchführung des Vorhabens benutzt wurden

Von Seiten der OVGU aus wurden keine externen Konstruktionen, Verfahren oder Schutzrechte verwendet. Die Forschungsarbeiten beruhten auf universitätsintern entwickelten Ansätzen und auf dem jeweils aktuellen Stand der Wissenschaft.

4.2 Verwendeten Fachliteratur

Ein Teil der Während der Projektlaufzeit genutzten und Veröffentlichten Fachliteratur kann der nachfolgenden Tabelle entnommen werden.

№	Quelle / Titel	Relevanz für Safe.trAIn	Verlag/ Konferenz
1	**Hawkins et al (2021):** Guidance on	Fundament für Safety Cases	Arxiv (2021)

№	Quelle / Titel	Relevanz für Safe.trAIn	Verlag/ Konferenz
	the Assurance of Machine Learning in Autonomous Systems (AMLAS)	ML-basierter Funktionen	
2	**VDE-AR-E 2842-61-2**: Anwendungsregel Development and trustworthiness of autonomous/cognitive systems	Normative Grundlage für Vertrauen in autonome Systeme	VDE
3	**ISO/IEC 23053**: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	Standard zur Einbettung von ML in AI-Systemprozesse	ISO/IEC
4	**EN 50126-1**: Railway Applications - The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) - Part 1: Generic RAMS Process	Relevanter Safety-Standard für Eisenbahnsysteme	CENELEC
5	**Hendrycks & Gimpel (2017)**: A baseline for detecting misclassified and out-of-distribution examples in neural networks	Grundlegende Methode zum OOD-Scoring	ICLR
6	**Lee et al. (2018)**: Mahalanobis-Distanz zur OOD-Erkennung	Statistisches Verfahren für Out-of-Distribution Detection	NeurIPS
7	**Domingos & Richardson (2006)**: Markov Logic Networks	Grundlage für probabilistische symbolische Systeme	ML Journal
8	**Zeller et al. (2024)**: Towards a Safe MLOps Process	Safe.trAIn: Methodische und technische MLOps-Integration	AI and Ethics
9	**Burton et al. (2019)**: Confidence Arguments für ML in Automated Driving	Argumentationsstruktur für Sicherheitsnachweise	SAFECOMP
10	**Sculley et al. (2015)**: Hidden Technical Debt in Machine Learning Systems	Motiviert Engineering-Aufwand in ML-Sicherheit	NeurIPS
11	**Dhamija et al. (2018)**: Reducing Network Agnostophobia	OOD-Detection im Kontext unsicherer Entscheidungen	NeurIPS
12	**Russakovsky et al. (2015)**: ImageNet	Referenz für Visual Perception Benchmarks	CVPR
13	**Kelly & Weaver (2004)**: The Goal Structuring Notation – A safety argumentation notation	Safety-Argumentation für ML-Komponenten	Dependable systems and networks workshop on assurance cases
14	**Borg et al. (2023)**: Safety Case für ML-Komponenten in der Automobilbranche	Anwendung der AMLAS-Prinzipien in der Praxis	Software Quality Journal

№	Quelle / Titel	Relevanz für Safe.trAIn	Verlag/ Konferenz
15	**Kirchheim et al. (2022):** PyTorch-OOD: A library for Out-of-Distribution Detection	Eigene Methodenentwicklungen so wie Weiterentwicklung im Safe.trAIn-Kontext	CVPRW
16	**Kirchheim et al. (2024)**: Out-of-Distributon Detection with Logical Reasonng	Eigene Methodenentwicklungen im Safe.trAIn-Kontext	WACV
17	**Kirchheim et al. (2025)**: Improvong Out-of-Distribution Detection with Markov Logic Networks	Eigene Methodenentwicklungen im Safe.trAIn-Kontext	ICML

II Abschlussbericht - Eingehende Darstellung

1 Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele & Notwendigkeit und Angemessenheit der geleisteten Arbeiten

Die Arbeiten der Otto-von-Guericke-Universität Magdeburg (OVGU) im Rahmen des Forschungsprojekts *„Safe.trAIn – Sichere KI am Beispiel fahrerloser Regionalzug“* verfolgten das Ziel, vertrauenswürdige Methoden der Künstlichen Intelligenz (KI) für sicherheitskritische Anwendungen im Bahnverkehr zu erforschen, zu entwickeln, zu evaluieren und systemisch zu integrieren. Die von der OVGU geleistete Arbeit erstreckte sich dabei über eine Vielzahl technischer, konzeptioneller und koordinativer Aufgaben, die in ihrer Gesamtheit einen wesentlichen Beitrag zum Erfolg des Projekts darstellten.

Im Folgenden werden die durchgeführten Arbeiten so wie die aus diesen resultierenden Ergebnisse des Projektes in den jeweiligen Arbeitspaketen vorgestellt und mit den geplanten Arbeiten verglichen.

1.1 Arbeitspaket 1 – Anforderungen an die Sicherheitsnachweisführung

Bereits zu Projektbeginn beteiligte sich die OVGU aktiv an den Arbeiten im AP1, welches der Analyse und Bewertung bestehender Normen und Standards zur sicherheitstechnischen Einordnung KI-basierter Systeme diente. Die Herausforderung bestand darin, die aus klassischen Domänen bekannten sicherheitstechnischen Regelwerke – etwa die EN 50129 für den Bahnbereich oder ISO 26262 für die Automobilbranche – hinsichtlich ihrer Übertragbarkeit auf nichtdeterministische, datengetriebene KI-Methoden zu evaluieren. Die OVGU übernahm hierbei eine unterstützende Rolle und arbeitete eng mit normenerfahrenen Projektpartnern wie dem TÜV Rheinland und der Siemens AG zusammen.

Insbesondere brachte die OVGU ihr Fachwissen in die systematische Analyse ein, um Unsicherheitsmetriken, Erklärbarkeit und Nachvollziehbarkeit in sicherheitsgerichtete Zertifizierungsprozesse zu integrieren. Diese Arbeiten bildeten eine essenzielle Grundlage für die spätere Argumentationsstruktur im Sicherheitsnachweis.

1.1.a) UAP 1.3 Anforderungen an Methoden, Werkzeuge, Entwicklungsprozesse, Betrieb

Im Rahmen von UAP 1.3 unterstützte die OVGU die Projektpartner bei der Ableitung spezifischer prozessualer Anforderungen für die Entwicklung und den Betrieb von Machine-Learning-Komponenten im sicherheitskritischen Bahnumfeld. Ausgehend von den in UAP 1.2 identifizierten Vertrauenseigenschaften – insbesondere

Erklärbarkeit, Robustheit und Unsicherheitsbehandlung – wurden bestehende Software-Engineering-Prozessmodelle analysiert und deren Übertragbarkeit auf KI-Entwicklungsprozesse untersucht. Berücksichtigt wurden KI-spezifische Aspekte wie regelmäßige Datenvalidierung, Sicherstellung der Datenqualität und der Umgang mit Modellunsicherheiten im Betrieb.
Gemeinsam mit den Partnern formulierte die OVGU Anforderungen, die eine kontinuierliche Prüfung, Dokumentation und Überwachung von KI-Komponenten über den gesamten Lebenszyklus gewährleisten. Damit wurde eine methodische Grundlage geschaffen, um klassische Sicherheitsstandards (z. B. EN 50128) mit KI-spezifischen Anforderungen zu verbinden und die Integration in zertifizierbare Architekturen zu ermöglichen. Die Ergebnisse flossen in nachgelagerte Arbeiten zur Verifikation (UAP 2.7), zum Monitoring (UAP 2.8) sowie in projektrelevante DIN SPECs ein.

Vorgegebenes Ziel	Erreichtes Ergebnis
Definition fundierter Anforderungen an Methoden, Werkzeuge sowie Entwicklungs- und Betriebsprozesse für KI-basierte Systeme im Bahnumfeld	Unterstützung bei Erstellung eines konsistenten Anforderungskatalogs, der klassische Sicherheitsstandards mit KI-spezifischen Prozessen verbindet
Unterstützung der Projektpartner bei der Ableitung prozessualer Anforderungen	Durchführung gemeinsamer Analyse- und Abstimmungsmeetings, Ergebnis: abgestimmte Prozessanforderungen für Entwicklung und Betrieb
Untersuchung der Übertragbarkeit klassischer Software-Engineering-Prozessmodelle auf KI-Entwicklung	Unterstützung bei der Identifikation notwendiger Anpassungen, insbesondere zu Datenvalidierung, Datenqualität und Unsicherheitsbehandlung
Schaffung einer Grundlage für Verifikation, Monitoring und Standardisierung	Erfolgreich durch Übergabe der Ergebnisse an UAP 2.3 (Validierung der Datengrundlage), 2.7 (Verifikation), 2.8 (Laufzeit-Monitoring) und an die Entwicklung der DIN SPECs.

1.1.b) UAP 1.4 Quantifizierbare Metriken zur Bewertung der Vertrauenswürdigkeit,

Im Rahmen von UAP 1.4 untersuchte die OVGU, wie sich Vertrauenswürdigkeit, Erklärbarkeit und sicherheitsrelevante Leistungsfähigkeit von KI-Systemen objektiv und quantifizierbar messen lassen. Die Ableitung und Bewertung geeigneter Metriken ist eine Grundvoraussetzung für den Sicherheitsnachweis KI-basierter Systeme im Bahnumfeld, insbesondere bei der Objekterkennung im autonomen Regionalzugbetrieb.

Dazu begleitete die OVGU die Recherche existierender Metriken und Bewertungsverfahren aus Forschung, Industrie und normativen Quellen. Aufbauend auf der in UAP 1.2

definierten Klassifikation vertrauenswürdiger Eigenschaften wurden Metriken systematisch analysiert, verglichen und auf ihre Eignung für sicherheitskritische ML-Anwendungen bewertet. Der Fokus lag u. a. auf Robustheit gegenüber Datenabweichungen, Vorhersageunsicherheit, Erklärbarkeit sowie Bias- und Fairness-Aspekten.

Darüber hinaus entwickelte die OVGU methodische Vorschläge für den Einsatz dieser Metriken entlang des gesamten Entwicklungszyklus – von der Datenanalyse über das Training und Testen bis zur Laufzeitüberwachung. Dabei wurden sowohl klassische Leistungskennzahlen (z. B. Genauigkeit, Precision/Recall, Intersection-over-Union) als auch fortgeschrittene Konzepte wie Confidence Calibration, OOD-Detection Scores und Datenabdeckungsmetriken berücksichtigt.

In Zusammenarbeit AP 3 und AP 4 wurden kontextspezifische Anforderungen für den autonomen Fahrbetrieb erarbeitet, insbesondere zu notwendiger Genauigkeit und Interpretierbarkeit für sicherheitstechnische Entscheidungen. Die Ergebnisse bildeten eine methodische Grundlage für spätere Sicherheitsnachweise, die Integration in Safety-Monitoring-Systeme, Verifikationsszenarien (AP 4) sowie projektspezifische Standardisierungsbeiträge. Teile der Arbeit sind in öffentlich zugängliche Softwarebibliotheken eingeflossen und wurden auf internationalen Konferenzen vorgestellt.

Vorgegebenes Ziel	Erreichtes Ergebnis
Entwicklung eines methodischen Rahmens zur Bewertung der Vertrauenswürdigkeit und Leistungsfähigkeit von KI-Systemen	Systematische Sammlung, Analyse und Bewertung relevanter Metriken für sicherheitskritische ML-Anwendungen
Berücksichtigung der in UAP 1.2 definierten Vertrauenseigenschaften	Unterstützung bei Integration von Robustheit, Erklärbarkeit, Unsicherheitsbehandlung und Fairness in die Metrikbewertung
Abstimmung kontextspezifischer Metrikanforderungen mit Partnern	Definition von Anfordrungen an Performance-Meriken für sicherheitstechnische Entscheidungen im autonomen Fahrbetrieb
Beitrag zu Sicherheitsarchitektur, Verifikationsszenarien und Standardisierung	Ergebnisse in Safety-Monitoring-Systeme, AP 4-Szenarien und DIN SPEC-Entwürfe integriert

1.2 Arbeitspaket 2 – Prüfmethoden und -werkzeuge zum Nachweis der Vertrauenswürdigkeit von KI-Methoden

Das zweite Arbeitspaket stellte den zentralen Arbeitsschwerpunkt der OVGU dar. Als wissenschaftliche Einrichtung war die OVGU nicht nur maßgeblich an inhaltlichen Beiträgen beteiligt, sondern übernahm auch die Leitung von zwei besonders kritischen Unterarbeitspaket: 2.3 und 2.7.

Zusätzlich entwickelte die OVGU Konzepte zur Kombination regelbasierter Sicherheitssysteme mit datengetriebenen KI-Methoden. Ziel war es, die interpretierbaren Eigenschaften klassischer Systeme mit der Leistungsfähigkeit moderner Modelle zu verknüpfen, um so die Transparenz und Verifikation zu stärken.

Die entwickelten Verfahren wurden nicht nur auf realistischen Datensätzen evaluiert, sondern in einem konsistenten Sicherheitsargumentationsbaum (GSN) verankert. Dadurch konnte die OVGU einen messbaren Beitrag zur Gesamtargumentation des Projekts leisten. Die Ergebnisse wurden projektintern vorgestellt und in mehreren internationalen Publikationen disseminiert. Eine der Veröffentlichungen wurde mit einem Best-Paper-Award ausgezeichnet, was die wissenschaftliche Relevanz der Arbeiten unterstreicht.

1.2.a) UAP 2.1 Analyse State-of-the-art in der Absicherung von KI-Funktionen

Die OVGU übernahm in UAP 2.1 die Aufgabe, den Stand von Forschung und Technik im Bereich der Absicherung von KI-Funktionen systematisch zu erfassen und im Hinblick auf sicherheitskritische Bahnanwendungen zu bewerten. Ziel war es, existierende Methoden zu sichten, ihre Anwendbarkeit im Bahnbereich zu klassifizieren und bestehende methodische Lücken zu identifizieren.

Dazu verfasste die OVGU einen umfassenden State-of-the-Art-Report zur Verifikation von KI-Systemen, der als internes Referenzdokument diente. Die Analyse umfasste ein breites Spektrum an Verfahren – von formalen Methoden über probabilistische Ansätze bis hin zu Testverfahren wie Coverage-Kriterien für neuronale Netze. Im Fokus stand deren Übertragbarkeit auf den Kontext des autonomen schienengebundenen Regionalverkehrs.

Besonderes Gewicht lag auf der Identifikation methodischer Defizite: Es fehlen standardisierte Verfahren zur Bewertung der Generalisierungsfähigkeit bei veränderten Umgebungsbedingungen (Out-of-Distribution-Szenarien) sowie formale Beschreibungen der Umweltbedingungen. Zudem erwies sich die Integration vieler bestehender Werkzeuge in sicherheitsgerichtete Entwicklungsprozesse als schwierig, da Traceability und Anbindung an Sicherheitsnachweisketten nicht hinreichend vorhanden sind.

Die Ergebnisse mündeten in eine Gap-Analyse mit konkreten Vorschlägen für nachgelagerte UAPs (insbesondere 2.3, 2.4 und 2.7). Damit legte die OVGU eine methodisch fundierte Grundlage für die Entwicklung neuer Metriken, die Erweiterung bestehender Werkzeuge und die Integration in den SafeMLOps-Lifecycle.

Vorgegebenes Ziel	Erreichtes Ergebnis
Systematische Analyse des Stands von Forschung und Technik zur Absicherung von KI-Funktionen	Erstellung eines umfassenden State-of-the-Art-Reports als internes Referenzdokument
Sichtung und Klassifikation existierender Methoden aus Forschung und Industrie	Kategorisierung relevanter Ansätze (formale Methoden, probabilistische Verfahren, Testverfahren) im Bahnumfeld

Vorgegebenes Ziel	Erreichtes Ergebnis
Untersuchung der Übertragbarkeit bestehender Verfahren und Tools auf den Projektkontext	Bewertung der Anwendbarkeit auf den autonomen Regionalverkehr, Identifikation von Anpassungsbedarf
Aufdeckung methodischer Lücken	Identifikation fehlender Ansätze zu Generalisierung, Out-of-Distribution-Absicherung und formaler Beschreibung von Umgebungsbedingungen
Unterstützung nachgelagerter Arbeitspakete durch strukturierte Ergebnisse	Durchführung einer Gap-Analyse mit konkreten Vorschlägen für UAP 2.3, 2.4 und 2.7

1.2.b) UAP 2.2 Konzept zur systematischen Erstellung eines Nachweises für die Vertrauenswürdigkeit der KI-Funktionen

In UAP 2.2 unterstützte die OVGU die Entwicklung eines ganzheitlichen Konzepts zur Sicherheitsnachweisführung für KI-basierte Perzeptionssysteme. Basierend auf den in AP 1 erarbeiteten Anforderungen wurde eine strukturierte Methodik entwickelt, um Eigenschaften wie Verlässlichkeit, Erklärbarkeit und Leistungsfähigkeit systematisch nachweisbar zu machen.

Ein wesentlicher Beitrag der OVGU lag in der Spezifikation geeigneter Metriken zur Quantifizierung sicherheitsrelevanter Eigenschaften. Dazu gehörten insbesondere Metriken zur Behandlung von Unsicherheiten, zur Out-of-Distribution-Erkennung und zur Segmentierungsqualität. Diese Metriken wurden in enger Zusammenarbeit mit dem Fraunhofer IKS in die Sicherheitsnachweisstruktur (GSN-Tree) integriert und als konkrete Evidenzen verankert, sodass ihr Beitrag zur Absicherung einzelner Systemfunktionen nachvollziehbar wurde.

Durch die Kombination quantitativer Nachweise mit strukturierten Argumentationsketten trug die OVGU entscheidend zur Operationalisierung des Vertrauenswürdigkeitsnachweises bei. Damit wurde eine methodische Grundlage für spätere Zulassungsdiskussionen sowie für die Bewertung durch externe Prüfstellen geschaffen.

1.2.c) UAP 2.3 Methodik zur Beschreibung der Betriebsumgebung und Sicherstellung der Datenqualität

In UAP 2.3 übernahm die OVGU eine koordinierende Rolle bei der Entwicklung eines methodischen Frameworks zur Beschreibung der Operational Design Domain (ODD) und zur Sicherstellung der Datenqualität. Ziel war es, eine präzise Modellierung des sicherheitsrelevanten Systemkontexts zu ermöglichen – sowohl als Grundlage für die Nachweisführung als auch für datengetriebene Entwicklungsprozesse.

Die OVGU brachte ihre Expertise in modellbasierter Sicherheitsanalyse im Bahnumfeld ein und leitete gemeinsam mit Fraunhofer IKS die Konzeption einer ontologischen Struktur zur Beschreibung des Systemkontexts. Ein Schwerpunkt lag auf der Verknüpfung von Umgebungskontext und Trainingsdaten: Die OVGU unterstützte das Fraunhofer IKS bei de rErarbeitung von Methoden, um Datensätze mit Kontextinformationen wie Wetter, Streckenabschnitt, Sichtbedingungen oder Sensorverfügbarkeit anzureichern. Damit wurde überprüfbar, ob Datensätze die gesamte relevante Domäne abdecken oder ob sicherheitskritische Lücken bestehen.

Die Arbeiten stellten sicher, dass KI-basierte Perzeptionssysteme nur in wohldefinierten und nachvollziehbar abgedeckten Betriebsumgebungen eingesetzt werden. Damit leistete die OVGU einen wesentlichen Beitrag für die spätere Integration in Verifikations-, Validierungs- und Monitoringprozesse.

Vorgegebenes Ziel	Erreichtes Ergebnis
Entwicklung eines methodischen Frameworks zur Beschreibung der Betriebsumgebung (ODD)	Konzeption einer semantischen Struktur zur präzisen Modellierung des Systemkontexts
Sicherstellung der Qualität der im System eingesetzten Daten	Entwicklung von Verfahren zur Anreicherung von Datensätzen mit Kontextinformationen (z. B. Wetter, Sichtbedingungen, Sensorverfügbarkeit)
Verknüpfung von Umgebungskontext und Trainingsdaten	Methodik zur Prüfung der Abdeckung relevanter Domänen und Identifikation sicherheitskritischer Lücken in Datensätzen
Beitrag zur Entwicklung robuster Datenqualitäts- und Abdeckungsmetriken	Grundlage geschaffen für weiterführende Arbeiten in nachgelagerten UAPs (z. B. Metriken zur ODD-Abdeckung)

1.2.d) UAP 2.5 Robustheit von KI-Funktionen

In UAP 2.5 beteiligte sich die OVGU an der Entwicklung robusterer KI-Verfahren mit dem Schwerpunkt auf Unsicherheitsmodellierung und kontextsensitiver Anomaliedetektion. Ziel war es, Methoden zu schaffen, die Perzeptionssysteme auch unter unsicheren oder veränderten Bedingungen absichern und das Risiko kritischer Fehlentscheidungen reduzieren.

Ein Fokus des Arbeitspaketes lag auf der Abschätzung und Kalibrierung von Modellunsicherheiten, die gezielt in sicherheitsgerichtete Entscheidungen eingebunden werden können. Zudem entwickelte die OVGU Ansätze zur Erkennung und Behandlung von Anomalien – etwa durch unbekannte Umwelteinflüsse oder nicht repräsentierte Eingaben – im Einklang mit den definierten Sicherheitsanforderungen.

Ein wesentliches Ergebnis war die Entwicklung eines neuartigen Trainingsverfahrens zur Out-of-Distribution (OOD)-Erkennung unter Nutzung adversarial generierter Ausreißer. Dieser Ansatz verbesserte nicht nur die Erkennungsleistung unbekannter Eingaben,

sondern steigerte auch die Robustheit gegenüber gezielten Störungen. Die Ergebnisse wurden im Rahmen einer wissenschaftlichen Publikation vorgestellt und auf dem CVPR-Workshop *Safe Artificial Intelligence for All Domains (SAIAD)* veröffentlicht, was den innovativen Beitrag der OVGU zur robusten KI-Absicherung unterstreicht.

Vorgegebenes Ziel	Erreichtes Ergebnis
Entwicklung robusterer KI-Verfahren für sicherheitskritische Anwendungen	Entwicklung und Validierung neuer Methoden zur Unsicherheitsmodellierung und Anomaliedetektion
Modellierung und Kalibrierung von Unsicherheit	Entwicklung von Verfahren zur quantitativen Abschätzung und Kalibrierung von Modellunsicherheiten für sicherheitsgerichtete Entscheidungen
Verbesserung der Robustheit gegenüber unsicheren und veränderten Bedingungen	Training mit adversarial generierten Ausreißern zur Steigerung der Erkennungsleistung und Robustheit
Dissemination wissenschaftlicher Ergebnisse	Veröffentlichung einer Publikation auf dem CVPR-Workshop SAIAD

1.2.e) UAP 2.6 Transparenz von KI-Funktionen

Die OVGU leistete in UAP 2.6 einen gezielten Beitrag zur Erhöhung der Transparenz und Erklärbarkeit von KI-Funktionen mit Schwerpunkt auf Out-of-Distribution (OOD) Detection. Ziel war es, Methoden zu entwickeln, die unbekannte Eingaben zuverlässig erkennen und deren Entscheidungsverhalten für Entwickler und Prüfer nachvollziehbar machen.

Im Fokus stand ein hybrider Ansatz, der Deep-Learning-Modelle mit formalen, regelbasierten Beschreibungen kombiniert. Die OVGU entwickelte einen neuartigen OOD-Detektor, der neben der datengetriebenen Vorverarbeitung eine wissensbasierte Komponente enthält: Ein logikbasiertes System überprüft zur Laufzeit, ob Eingaben mit dem bekannten Strukturwissen über die Trainingsdomäne konsistent sind. So konnte die Erkennungsleistung in sicherheitskritischen Grenzfällen verbessert und die Erklärbarkeit der Entscheidungen deutlich erhöht werden.

Im Unterschied zu klassischen, rein datengetriebenen Verfahren erlaubt der Ansatz der OVGU eine explizite Steuerung und Erweiterung des Systemverhaltens durch menschlich lesbare Regeln. Diese semantische Nachvollziehbarkeit ist insbesondere für sicherheitskritische Anwendungen relevant, in denen Vertrauen in die Reaktionslogik essenziell ist.

Die Ergebnisse wurden in einer wissenschaftlichen Publikation auf einer internationalen Konferenz vorgestellt und zeigen exemplarisch, wie erklärbare KI-Verfahren sowohl technische als auch regulatorische Anforderungen erfüllen können. Damit schlug die OVGU die Brücke zwischen moderner KI-Forschung und den Anforderungen sicherheitsgerichteter Entwicklung.

Vorgegebenes Ziel	Erreichtes Ergebnis
Entwicklung von Methoden zur zuverlässigen Erkennung unbekannter Eingaben	Neu entwickelter hybrider OOD-Detektor mit kombinierter datengetriebener und wissensbasierter Komponente
Sicherstellung von Transparenz und Erklärbarkeit der Detektionslogik	Einführung eines logikbasierten Systems zur Laufzeitprüfung der Eingabekonsistenz
Verbesserung der Leistungsfähigkeit in sicherheitsrelevanten Grenzfällen	Nachweis erhöhter Erkennungsleistung bei gleichzeitiger Erklärbarkeit
Dissemination wissenschaftlicher Ergebnisse	Veröffentlichung einer Publikation auf internationaler Konferenz zu erklärbarer OOD Detection

1.2.f) UAP 2.7 Verifikation von KI-Funktion

Die OVGU übernahm in UAP 2.7 die Leitung und koordinierte die Entwicklung von Verfahren zur Verifikation sicherheitskritischer KI-Funktionen. Ziel war es, belastbare methodische Ansätze zu schaffen, mit denen das intendierte Verhalten von KI-Systemen überprüfbar und nachvollziehbar nachgewiesen werden kann – eine essenzielle Voraussetzung für deren Einsatz im autonomen Schienenverkehr.

Ein Schwerpunkt lag auf der Entwicklung quantitativer Metriken zur Leistungsbewertung sicherheitsrelevanter Komponenten, wie etwa der Objektsegmentierung. Aufbauend auf den Ergebnissen aus UAP 1.4 wurden Bewertungsverfahren entwickelt, die Zuverlässigkeit und Robustheit differenziert erfassen – auch bei wechselnden Umgebungsbedingungen, veränderten Eingabestrukturen und Unsicherheiten.

Darüber hinaus erarbeitete die OVGU Methoden zur strukturierten Partitionierung des Eingaberaums. Ziel war es, kritische und schwer vorhersagbare Bereiche zu identifizieren und gezielt zu testen, um auch seltene Randfallszenarien in die Verifikation einzubeziehen (vgl. UAP 2.3). Damit wurde eine intelligente Ergänzung zu klassischen testbasierten Verfahren geschaffen, bei denen vollständige Abdeckung nicht realistisch möglich ist.

Als Leiterin des UAP koordinierte die OVGU zudem die Zusammenarbeit zwischen verschiedenen Partnern. Die entwickelten Metriken und Analysemethoden wurden schließlich in die strukturierte Sicherheitsargumentation (GSN) integriert und bilden heute einen zentralen Bestandteil der projektierten Safety-Architektur.

Vorgegebenes Ziel	Erreichtes Ergebnis
Entwicklung belastbarer Verifikationsverfahren für KI-Systeme	Konzeption und Umsetzung von Methoden zur Verifikation sicherheitskritischer KI-Komponenten
Quantitative Bewertung der Leistungsfähigkeit von KI-Funktionen	Entwicklung von Metriken für Zuverlässigkeit, z. B. bei Objektsegmentierung
Strukturierte Partitionierung des Eingaberaums zur Identifikation kritischer Szenarien	Entwicklung von Methoden zur systematischen Erfassung und Testung von Randfallszenarien, Ergänzung zu testbasierter Verifikation

Vorgegebenes Ziel	Erreichtes Ergebnis
Koordination der Partner zur Entwicklung einer einheitlichen Verifikationsstrategie	Leitung des UAP durch OVGU, Organisation der Zusammenarbeit von methodischen, daten- und sicherheitsgetriebenen Partnern
Integration der Ergebnisse in den Sicherheitsnachweis	Einbindung der entwickelten Methoden in die GSN-Argumentation als zentraler Baustein der Safety-Architektur

1.2.g) UAP 2.8 Laufzeit-Maßnahmen zur Sicherstellung der KI-Verlässlichkeit

Die OVGU beteiligte sich in UAP 2.8 an der Entwicklung von Laufzeit-Monitoringverfahren, die darauf abzielen, die Verlässlichkeit von KI-Systemen im Betrieb kontinuierlich zu bewerten und Fehlfunktionen frühzeitig zu erkennen. Der Fokus lag auf der Out-of-Distribution (OOD) Detection – der Identifikation von Eingaben, die außerhalb der Trainingsverteilung liegen und damit ein erhöhtes Risiko für unerwartetes Verhalten bergen.

Dazu entwickelte die OVGU verschiedene Metriken zur Laufzeitbewertung der OOD-Erkennung, die eine Quantifizierung der Detektionsleistung und Rückschlüsse auf die Zuverlässigkeit des Gesamtsystems ermöglichen. Dabei wurden sowohl einfache statistische Verfahren als auch komplexere Unsicherheitsindikatoren betrachtet. Parallel wurde die Performanz der Verfahren analysiert, um ihre Echtzeitfähigkeit sicherzustellen.

Ein zentrales Ergebnis war die Entwicklung eines hybriden OOD-Detektionsansatzes auf Basis logisch-probabilistischer Modelle. Dieser kombiniert klassische OOD-Detektoren mit Markov Logic Networks (MLNs), die probabilistische Inferenz mit symbolischem Wissen verknüpfen. Dadurch entstand ein Verfahren, das robust und erklärbar zugleich ist. Ergänzend wurde ein Algorithmus entwickelt, der logische Regeln aus Trainingsdaten extrahiert, um die Modellkonsistenz während der Laufzeit abzusichern.

Die Ergebnisse wurden in einer wissenschaftlichen Publikation zusammengefasst, welche bei der ICML akzeptiert und veröffentlicht wurde. Damit leistete die OVGU einen wesentlichen Beitrag zur Integration von OOD Detection in die Betriebsüberwachung sicherheitskritischer Systeme. Die Verfahren erweitern das Spektrum der Evidenzen im Sicherheitsnachweis um dynamische, datengetriebene Komponenten.

Vorgegebenes Ziel	Erreichtes Ergebnis
Entwicklung von Laufzeit-Monitoringverfahren für KI-Systeme	Entwicklung und Erprobung von Verfahren zur kontinuierlichen Bewertung der Verlässlichkeit im Betrieb
Entwicklung von Metriken zur Laufzeitbewertung	Konzeption von statistischen und komplexeren Unsicherheitsmetriken zur Quantifizierung der Detektionsleistung
Sicherstellung der Echtzeitfähigkeit	Evaluation der Performanzanforderungen,

Vorgegebenes Ziel	Erreichtes Ergebnis
	Nachweis der Integrationstauglichkeit in den Betrieb
Nutzung symbolischer Modelle zur Laufzeitabsicherung	Algorithmus zur automatischen Ableitung logischer Regeln aus Trainingsdaten zur Konsistenzprüfung
Dissemination wissenschaftlicher Ergebnisse	Veröffentlichung einer Publikation zur OOD Detection mit MLNs (angenommen bei ICML nach Projektende)

1.3 Arbeitspaket 3 – Fahrzeugarchitektur im GoA4-Betrieb mit dem Fokus auf sichere KI-basierte Funktionen

In AP3 arbeitete die OVGU an der konzeptionellen und technischen Einbettung sicherer KI-Verfahren in die Architektur eines fahrerlosen Regionalzugs. Insbesondere konzentrierte sie sich darauf, die in AP2 entwickelten Verfahren – etwa OOD-Erkennung, Segmentierungsmetriken oder Unsicherheitsquantifizierung – in ein redundantes, fehlertolerantes Gesamtsystem zu integrieren.

Hierzu wurden Vorschläge für alternative Verifikationspfade und Fallback-Mechanismen gemacht, um die Systemstabilität bei unsicheren KI-Vorhersagen zu sichern. Die OVGU war in regelmäßige Architekturtreffen mit Partnern wie Siemens, Fraunhofer IKS und BIT eingebunden und stellte sicher, dass Sicherheitsaspekte frühzeitig in die Systementwicklung einflossen. Diese Abstimmungsarbeit war aufwendig, aber notwendig, um die späteren Schnittstellen zur Implementierung (AP4) konsistent vorzubereiten.

1.3.a) UAP 3.2 Spezifikation Operational Design Domain (ODD)

In UAP 3.2 unterstützte die OVGU die Projektpartner bei der Spezifikation der ODD für den fahrerlosen Regionalzug. Aufbauend auf den in UAP 2.3 entwickelten Kontextmodellen lag der Schwerpunkt auf der systematischen Definition der ODD-Grenzen sowie auf der Identifikation dynamischer und uneindeutiger Grenzbereiche, die für die Sicherheit besonders kritisch sind.

Ziel war es, ein konsistentes, strukturiertes ODD-Modell zu entwickeln, das Umwelt-, Infrastruktur- und Interaktionsfaktoren des Bahnbetriebs integriert. Die OVGU brachte dabei ihre Expertise aus dem Bereich schienengebundener Verkehr und modellbasierter Sicherheitsanalyse ein.

Ein besonderer Fokus lag auf der Definition von ODD-Grenzverletzungen: Es wurden Kriterien erarbeitet, mit denen das System erkennen kann, wenn es sich der ODD-Grenze nähert oder diese überschreitet. Solche Mechanismen sind essenziell, um zur Laufzeit sicherheitsgerichtete Maßnahmen wie regelbasierte OOD Detection zu aktivieren.

Durch die Beteiligung der OVGU konnten technische und sicherheitsrelevante Aspekte in ein praxistaugliches Modell überführt werden, das sowohl für Validierung als auch für den Sicherheitsnachweis genutzt werden kann.

Vorgegebenes Ziel	Erreichtes Ergebnis
Erstellung eines konsistenten, strukturierten ODD-Modells	Entwicklung eines Modells, das Umwelt-, Infrastruktur- und Interaktionsfaktoren integriert (entsprechend der in AP 2 entwickelten Methoden)
Systematische Definition von ODD-Grenzen	Formulierung klarer ODD-Grenzen für den fahrerlosen Regionalzug
Unterstützung von Validierung und Sicherheitsnachweis	Bereitstellung eines praxistauglichen ODD-Modells für spätere Verifikations- und Nachweisprozesse

1.3.b) UAP 3.5 Architektur Sicheres Objekterkennungssystem

Die OVGU beteiligte sich in UAP 3.5 an der Konzeption der Systemarchitektur eines sicheren Objekterkennungssystems. Zentrale Aufgabe war es, die Ergebnisse aus AP2 – insbesondere OOD Detection, Unsicherheitsmetriken und Datenqualitätsbewertung – in eine konsistente und praktisch umsetzbare Architektur zu überführen.

Ein Schwerpunkt lag auf der Integration von OOD-Mechanismen zur Erkennung großer, potenziell unbekannter Objekte außerhalb der Trainingsverteilung. Diese Erweiterung klassischer Objekterkennung ist im Bahnumfeld entscheidend, da hier mit unvorhersehbaren Szenarien wie Gegenständen oder Personen auf den Gleisen gerechnet werden muss. Die OVGU untersuchte mögliche Systempfade, um die OOD-Erkennung mit bestehenden Perzeptionsmodulen zu koppeln und so eine robuste Entscheidungsgrundlage auch in unsicheren Situationen zu schaffen.

Darüber hinaus wirkte die OVGU an der Gestaltung der Testdatenerzeugung mit. Sie definierte Anforderungen an Szenariokonfigurationen, die sicherheitskritische Situationen gezielt simulieren, insbesondere Randfälle und seltene Objekte. Damit konnte die Validität der entwickelten Metriken unter realitätsnahen Bedingungen in einem sicheren Umfeld überprüft werden.

Insgesamt stellte die OVGU sicher, dass sicherheitsgerichtete KI-Komponenten nicht isoliert, sondern systematisch mit allen relevanten Datenflüssen in die Gesamtarchitektur integriert werden. Damit leistete sie einen notwendigen Beitrag, um das Objekterkennungssystem auf eine solide sicherheitstechnische Grundlage zu stellen.

Vorgegebenes Ziel	Erreichtes Ergebnis
Integration der Ergebnisse aus AP2 (OOD, Unsicherheitsmetriken, Datenqualität) in die Systemarchitektur	Entwicklung einer konsistenten Architektur mit systematischer Einbettung von KI-Sicherheitsmechanismen
Absicherung gegen unbekannte und unvorhersehbare Szenarien	Mechanismen zur Erkennung großer, unbekannter Objekte außerhalb der

Vorgegebenes Ziel	Erreichtes Ergebnis
	Trainingsverteilung
Mitgestaltung der Testdatenerzeugung für sicherheitskritische Szenarien	Definition von Anforderungen und Beispielszenarien zur Simulation von Randfällen und seltenen Objekten
Gewährleistung einer belastbaren sicherheitstechnischen Basis	Systematische Verknüpfung sicherheitsgerichteter KI-Komponenten mit Architektur und Datenflüssen

1.3.c) UAP 3.6 Architektur Sichere Sensorfusion und Umgebungsmodell

In UAP 3.6 war die OVGU nicht direkt in die Entwicklung eigener Komponenten eingebunden. Sie begleitete jedoch die Arbeiten zur Architektur einer sicheren Sensorfusion, um den Informationsfluss und die Schnittstellen zu anderen sicherheitskritischen Systemteilen – insbesondere zur OOD Detection und zur Betriebsumgebungsmodellierung – nachvollziehen und gegebenenfalls abgestimmt beeinflussen zu können.

Die Teilnahme diente in erster Linie dazu, den Anschluss an die Systemarchitektur zu sichern und potenzielle Auswirkungen auf die Integration von KI-basierten Sicherheitsmaßnahmen frühzeitig zu erkennen. Ein eigener methodischer Beitrag erfolgte in diesem Arbeitspaket nicht.

Vorgegebenes Ziel	Erreichtes Ergebnis
Entwicklung einer sicheren Architektur für Sensorfusion	OVGU begleitete die Arbeiten und stellte Anschlussfähigkeit zur Gesamtarchitektur sicher
Abstimmung mit sicherheitskritischen Komponenten (z. B. OOD Detection, ODD-Modellierung)	Beobachtung und Einflussnahme auf Schnittstellen und Informationsflüsse

1.3.d) UAP 3.9 Entwicklung Sicheres Objekterkennungssystem und Sichere Sensorfusion

In UAP 3.9 unterstützte die OVGU die Implementierung und Integration von sicherheitsgerichteten KI-Methoden in das Objekterkennungssystem des Projekts. Aufbauend auf den zuvor entwickelten Konzepten und Metriken – insbesondere OOD Detection, Unsicherheitsabschätzung und Kontextbewertung – leistete die OVGU wesentliche Beiträge zur Überführung dieser Verfahren in funktionsfähige Softwarekomponenten.

Ein Schwerpunkt lag auf der Integration der entwickelten Metriken in bestehende Test-Pipelines, um sicherheitskritische Situationen zuverlässig zu erkennen und zu bewerten. In

enger Abstimmung mit den Systempartnern wurden Implementierungsdetails, Datenformate und Anforderungen abgestimmt, um eine reibungslose Anbindung an die Sensorik des Virtuellen Testsystems zu gewährleisten.

Darüber hinaus beteiligte sich die OVGU an der Definition und Umsetzung von Testszenarien, die sicherheitsrelevante Randfälle – etwa unerwartete Objekte auf der Strecke oder unvollständige Sensordaten – gezielt simulierten. Hierfür entwickelte sie spezifische Auswertungsmodule zur Berechnung der relevanten Metriken und führte erste Validierungen durch, um die Praxistauglichkeit der Komponenten sicherzustellen.

So wurde gewährleistet, dass die von der OVGU entwickelten Sicherheitsverfahren nicht nur konzeptionell fundiert, sondern auch praktisch einsetzbar und systemkompatibel sind – ein notwendiger Schritt zur Integration in das virtuelle Testfeld und zur Vorbereitung des sicherheitsgerichteten Gesamtnachweises.

Vorgegebenes Ziel	Erreichtes Ergebnis
Integration sicherheitsgerichteter KI-Methoden in das Objekterkennungssystem	Überführung von OOD Detection, Unsicherheitsmetriken und Kontextbewertung in funktionsfähige Softwarekomponenten
Einbindung der entwickelten Verfahren in Test-Pipelines	Technische Integration und Anbindung an Sensorik des Virtuellen Testsystems
Definition und Umsetzung von Testszenarien für sicherheitskritische Randfälle	Entwicklung von Szenarien für unerwartete Objekte und unvollständige Sensordaten
Entwicklung geeigneter Auswertungsmodule	Implementierung von Modulen zur Berechnung und Validierung relevanter Metriken
Sicherstellung der Praxistauglichkeit und Systemkompatibilität	Validierungen erfolgreich durchgeführt, Vorbereitung für Einbindung in Sicherheitsnachweis erfolgt

1.4 4. Arbeitspaket 4 – Virtuelles Testfeld, Sicherheitsbewertung

Die Aufgabe der OVGU in AP4 bestand im Wissenstransfer: Theoretische Forschungsergebnisse sollten unter realitätsnahen Bedingungen in einer virtuellen Testumgebung, so wie auch Realwelt-Daten validiert werden. Die Universität entwickelte hierfür eine Softwarebibliothek, welche die digitale ODD-Beschreibung interpretierbar macht und mit den zuvor entwickelten Metriken verknüpft.

Zusätzlich wurde die Integration von KI-Metriken in die Compute-Infrastruktur von Siemens umgesetzt. Diese Integration stellte eine große Herausforderung dar, da unterschiedliche Schnittstellen und Datenformate berücksichtigt werden mussten. In mehreren Iterationen wurden die Bibliotheken angepasst, erweitert und projektintern zur Verfügung gestellt. Rückmeldungen aus der Testfeldpraxis flossen direkt in die Weiterentwicklung ein. Die

OVGU konnte zeigen, dass ihre Verfahren nicht nur theoretisch fundiert, sondern auch praxisnah und einsatzfähig sind.

1.4.a) UAP 4.2 Konzeption virtuelles Testfeld

In UAP 4.2 unterstützte die OVGU die Konzeption und Anbindung des virtuellen Testfelds, das als simulationsbasierte Evaluierungsumgebung für die entwickelten KI-Funktionen diente. Der Schwerpunkt lag auf der Sicherstellung der Schnittstellenkompatibilität zwischen Testszenarien, Bewertungsmethoden und Testfeldarchitektur.

Die OVGU wirkte insbesondere an der Integration der zuvor entwickelten Metriken – etwa zur Unsicherheit, Out-of-Distribution (OOD)-Erkennung und ODD-Abdeckung – in die Simulationsumgebung mit. Sie prüfte deren technische Umsetzbarkeit und schlug bei Bedarf Anpassungen vor, um eine zuverlässige Bewertung sicherheitskritischer Szenarien zu gewährleisten.

Durch diese Arbeiten trug die OVGU wesentlich dazu bei, dass das virtuelle Testfeld für die Validierung der projektrelevanten Sicherheitsanforderungen geeignet ist und als Grundlage für den späteren Sicherheitsnachweis zur Verlässlichkeit von KI-Komponenten dienen kann.

Vorgegebenes Ziel	Erreichtes Ergebnis
Konzeption und Anbindung eines virtuellen Testfelds	Unterstützung bei Aufbau und Architektur des simulationsbasierten Testfelds
Sicherstellung der Schnittstellenkompatibilität zwischen Szenarien, Methoden und Architektur	Prüfung und Abstimmung der Schnittstellen, Sicherstellung reibungsloser Integration
Integration entwickelter Metriken (Unsicherheit, OOD, ODD-Abdeckung) in die Testumgebung	Iterative Einbindung der Verfahren in die Simulation, inkl. technischer Validierung
Anpassung der Testumgebung für sichere Szenariobewertung	Vorschläge und Umsetzungsschritte zur Optimierung der Bewertungsprozesse

1.4.b) UAP 4.3 Prototypische Umsetzung virtuelles Testfeld

In UAP 4.3 unterstützte die OVGU die prototypische Umsetzung des virtuellen Testfelds durch die Integration ihrer Softwarekomponenten in die gemeinsam genutzte Infrastruktur, insbesondere den *ai.store*. Ziel war es, Zugriff auf das Testfeld sowie die dort erzeugten Test- und Beispieldaten zu erhalten und diese für eigene Auswertungen und Sicherheitstests nutzbar zu machen.

Im Zuge der Anbindung nahm die OVGU gezielte Softwareanpassungen vor – etwa an den eigenen OOD-Erkennungsmodulen, den integrierten Sicherheitsmetriken und den Testszenarien. Dadurch konnte sichergestellt werden, dass die entwickelten Verfahren

reibungslos im Testfeldablauf nutzbar sind und mit der bestehenden Infrastruktur interoperieren.

Die Arbeiten der OVGU trugen dazu bei, das virtuelle Testfeld in einen funktionsfähigen Gesamtprototypen zu überführen, der als Basis für spätere, systematische Absicherungs- und Validierungsschritte dient.

Vorgegebenes Ziel	Erreichtes Ergebnis
Prototypische Umsetzung des virtuellen Testfelds	Integration der OVGU-Komponenten in die gemeinsame Testfeld-Infrastruktur
Zugriff auf Test- und Beispieldaten	Nutzung der im Testfeld generierten Daten für eigene Auswertungen und Sicherheitstests
Anbindung und Anpassung eigener Softwaremodule	Modifikationen an OOD-Erkennung, Sicherheitsmetriken und Testszenarien zur Interoperabilität
Sicherstellung der Funktionsfähigkeit im Gesamtsystem	Reibungslose Integration der Verfahren in den Testfeldablauf
Beitrag zur Überführung in einen Gesamtprototypen	Mitwirkung an der Etablierung eines funktionsfähigen Testfelds als Grundlage für Validierungsschritte

1.4.c) UAP 4.4 Testszenarien

In UAP 4.4 unterstützte die OVGU die Entwicklung von Testszenarien, die notwendig sind, um sicherheitsrelevante Anforderungen gezielt abzudecken und zu bewerten. Ziel war es, über das einfache Abfahren realer Strecken hinauszugehen und auch seltene, aber kritische Situationen realistisch zu erfassen.

Die OVGU arbeitete an der Konzeption dynamischer und statischer Szenarien, darunter ungewöhnliche Objekte auf der Strecke, Interaktionen mit anderen Verkehrsteilnehmern sowie verschiedene Witterungseinflüsse. Dabei wurde darauf geachtet, dass diese Szenarien die Anwendung der entwickelten Sicherheitsmetriken ermöglichen und statistisch aussagekräftige Ergebnisse liefern.

Zusätzlich wurden die Szenarien in die vorhandene Softwareumgebung integriert, sodass sie automatisiert innerhalb des virtuellen Testfelds abgerufen und ausgewertet werden können. Damit stellte die OVGU sicher, dass Anforderungen aus den vorhergehenden APs – insbesondere zu Datenqualität, ODD-Abdeckung und OOD Detection – auch praktisch überprüfbar sind.

Vorgegebenes Ziel	Erreichtes Ergebnis
Ermöglichung der Anwendung entwickelter Sicherheitsmetriken	Szenarien so gestaltet, dass Metriken zu Datenqualität, ODD-Abdeckung und OOD Detection nutzbar sind
Integration in die Softwareumgebung	Automatisierbare Testszenarien innerhalb des

Vorgegebenes Ziel	Erreichtes Ergebnis
des Testfelds	virtuellen Testfelds implementiert
Praktische Überprüfbarkeit theoretischer Konzepte aus vorherigen APs	Nachweis, dass Metriken und Modelle im operativen Testfeld evaluiert werden können

1.4.d) UAP 4.6: Evaluierung des sicheren Objekterkennungssystems und der KI-Methoden

In UAP 4.6 beteiligte sich die OVGU an der systematischen Evaluierung der entwickelten Sicherheitsmetriken und KI-Methoden innerhalb des virtuellen Testfelds. Ziel war es, die Effektivität der Verfahren zur OOD Detection, Unsicherheitsabschätzung und Datenqualitätsbewertung unter realitätsnahen Bedingungen zu prüfen und ihre Eignung für die sicherheitsgerichtete Nachweisführung zu validieren.

Hierfür führte die OVGU Tests in simulierten als als Realwelt-Szenarien durch, bei denen die Metriken berechnet und mit den jeweiligen Systemzuständen in Beziehung gesetzt wurden. Ein besonderer Fokus lag auf der Evaluation der Aussagekraft der Metriken: In enger Abstimmung mit den Partnern wurde diskutiert, wie gut diese Eigenschaften wie Generalisierungsfähigkeit oder ODD-Detection quantitativ belegen können. Die Ergebnisse wurden in sogenannten Metrik-Factsheets dokumentiert.

Alle berechneten Metriken und Evaluierungsergebnisse wurden im *ai.store* abgelegt, ergänzt um die jeweiligen Modell- und Datensatzversionen.

Mit diesen Arbeiten stellte die OVGU sicher, dass der Nutzen der entwickelten Methoden für die Sicherheitsnachweisführung nicht nur theoretisch begründet, sondern auch praktisch demonstriert werden konnte.

1.4.e) UAP 4.7 Beitrag zur Sicherheitsnachweisführung für das sichere Objekterkennungssystem

In UAP 4.7 unterstützte die OVGU die Projektpartner bei der exemplarischen Sicherheitsnachweisführung für das KI-basierte Objekterkennungssystem im Kontext fahrerloser Regionalzüge. Ziel war es, die im virtuellen Testfeld ermittelten Evidenzen – insbesondere Metriken und Testergebnisse – systematisch in die übergeordnete Sicherheitsargumentation zu integrieren.

Die OVGU trug dazu bei, Nachweise auf Grundlage der im *ai.store* abgelegten Metriken zu formulieren und mit den im Projekt definierten Prüfkriterien (aus AP1) zu verknüpfen. Eingebunden wurden vor allem jene Metriken, die in den UAPs zur OOD Detection, Unsicherheitsabschätzung und ODD-Abdeckung entwickelt und validiert worden waren.

Ein besonderer Fokus lag auf der Integration der Metriken in die Safety-Case-Struktur, die in der Plattform *nLoop* umgesetzt wurde. Die OVGU unterstützte sowohl konzeptionell –

durch die Formulierung nachvollziehbarer Nachweisargumente – als auch technisch, durch die Abstimmung der Datenflüsse zwischen Testumgebung, *ai.store* und Nachweisplattform.

Damit leistete die OVGU einen notwendigen Beitrag zur operativen Umsetzbarkeit des Sicherheitsnachweises, indem sie dabei half, konkrete Evidenzen mit Akzeptanzkriterien zu verknüpfen um zu demonstrieren, wie KI-bezogene Sicherheitsanforderungen nachvollziehbar erfüllt und dokumentiert werden können.

Vorgegebenes Ziel	Erreichtes Ergebnis
Exemplarische Sicherheitsnachweisführung für KI-basierte Objekterkennung	Unterstützung bei der Integration von Evidenzen in die Sicherheitsargumentation
Nutzung der Evidenzen aus dem virtuellen Testfeld	Formulierung von Nachweisen basierend auf im *ai.store* abgelegten Metriken und Testergebnissen
Verknüpfung mit definierten Prüfkriterien aus AP1	Systematische Einbindung relevanter Metriken (OOD Detection, Unsicherheit, ODD-Abdeckung)
Integration in Safety-Case-Struktur	Umsetzung in der Plattform *nLoop* mit konzeptioneller und technischer Unterstützung durch OVGU

1.5 Arbeitspaket 5 – Standardisierung und Verbreitung

Als wissenschaftliche Einrichtung war der OVGU die langfristige Verwertung im Bereich Forschung und Lehre ein besonderes Anliegen. Die entwickelten Konzepte flossen in Seminare, Abschlussarbeiten und Forschungsprojekte ein. Mehrere studentische Arbeiten wurden im Safe.trAIn-Kontext durchgeführt. Darüber hinaus war die OVGU aktiv in der Entwicklung der beiden DIN SPECs 99002 und 99004 eingebunden, welche zu den wichtigsten standardisierungsbezogenen Ergebnissen des Projekts zählen.

Auch auf internationalen Konferenzen war die OVGU präsent: Durch Vorträge auf WACV, WAISE/SafeComp, KI2023, ICML und weiteren Tagungen wurde das Projekt sichtbar gemacht, wissenschaftlich validiert und mit internationalen Partnern diskutiert. Diese Disseminationsarbeit war mit erheblichem organisatorischem und inhaltlichem Aufwand verbunden – etwa für die Vorbereitung von Vorträgen, die Veröffentlichung wissenschaftlicher Artikel oder die Teilnahme an Gremiensitzungen.

1.5.a) UAP 5.1: Identifikation von Normungs- und Standardisierungsbedarfen/-potentialen

In UAP 5.1 beteiligte sich die OVGU an der Analyse von Normungs- und Standardisierungsbedarfen für sichere KI im Bahnumfeld. Ziel war es, die Projektergebnisse frühzeitig auf ihre Anschlussfähigkeit an bestehende Regelwerke zu prüfen und mögliche Beiträge zu zukünftigen Standards zu identifizieren.

Vorgegebenes Ziel	Erreichtes Ergebnis
Prüfung der Anschlussfähigkeit der Projektergebnisse	Bewertung der Eignung von Methoden wie OOD Detection und Unsicherheitsmetriken
Ableitung möglicher Beiträge zu standardisierungsrelevanten Dokumenten	Benennung von Erweiterungs- und Anpassungsbedarfen
Unterstützung der Einbindung in Standardisierung	Frühzeitige Positionierung der Projektergebnisse im Normungsdiskurs

1.5.b) UAP 5.2 Bildung von Anwenderkreisen/Übertragung auf verwandte Use Cases in anderen Anwendungsdomänen der Sicherheitsargumentation

In UAP 5.2 brachte sich die OVGU in den Austausch mit internen und externen Stakeholdern ein, um die Generalisierbarkeit der entwickelten Methoden und Metriken über den Bahnbereich hinaus zu prüfen. Ein besonderer Fokus lag auf der Übertragbarkeit in die Robotik, wo ähnliche sicherheitskritische Herausforderungen – etwa unsichere Umgebungen, sensorbasierte Wahrnehmung und OOD-Signale – bestehen. Dadurch wurde die Anwendbarkeit der Projektergebnisse über den Use Case „fahrerloser Regionalzug“ hinaus demonstriert und neue Anknüpfungspunkte für zukünftige Forschung geschaffen.

Vorgegebenes Ziel	Erreichtes Ergebnis
Austausch mit Stakeholdern zur Generalisierbarkeit der Methoden	Teilnahme an projektinternen und externen Diskussionen
Prüfung der Übertragbarkeit auf andere Anwendungsbereiche	Analyse von Einsatzmöglichkeiten in der Robotik
Identifikation von Ähnlichkeiten in sicherheitskritischen Herausforderungen	Übertragung der Konzepte auf Themen wie Unsicherheit, OOD-Signale und Sensorfusion
Schaffung von Anknüpfungspunkten für zukünftige Forschung	Ableitung konkreter Perspektiven für KI-Safety in verwandten Domänen

1.5.c) UAP 5.3 Transfer der methodischen Vorgehensweise auf andere Anforderungen der Vertrauenswürdigkeit sowie deren Anwendungsbereiche

In UAP 5.3 untersuchte die OVGU, ob sich die im Projekt entwickelten Methoden – insbesondere Sicherheitsmetriken und das ODD-Framework aus AP2 und AP3 – auf weitere Dimensionen der Vertrauenswürdigkeit von KI-Systemen, wie Datenschutz oder Fairness, übertragen lassen. Eine konkrete, unmittelbare methodische Übertragbarkeit konnte noch nicht belegt werden. Die Arbeiten lieferten jedoch erste Impulse für künftige Forschung zur Erweiterung der Methoden über den Schienenverkehr hinaus.

Vorgegebenes Ziel	Erreichtes Ergebnis
Untersuchung der Übertragbarkeit entwickelter Methoden auf andere Vertrauenswürdigkeitsdimensionen	Analyse der Anwendbarkeit von Sicherheitsmetriken und ODD-Framework auf anderen Anwendungsfälle
Methodische Prüfung der Konformität jenseits der Sicherheit	Keine unkittelbare Übertragbarkeit belegbar

1.5.d) UAP 5.4 Initiierung und Umsetzung von Standardisierungsaktivitäten

In UAP 5.4 beteiligte sich die OVGU aktiv an der Überführung zentraler Projektergebnisse in normativ verwertbare Dokumente. Aufbauend auf den in UAP 1.1 sowie 5.1 bis 5.3 identifizierten Themenfeldern leistete sie Beiträge zu methodischen Grundlagen, Begriffsdefinitionen und technischen Anforderungen – insbesondere zu Sicherheitsmetriken, ODD-Beschreibung und der Absicherung KI-basierter Perzeptionssysteme.

Als zentrales Ergebnis wirkte die OVGU maßgeblich an zwei DIN SPECs mit:

- **DIN DKE SPEC 99002:2025 – Terminology – AI in railway applications**: Definition zentraler Begriffe und Konzepte zur Anwendung von KI im Bahnumfeld als gemeinsame Grundlage für zukünftige Regelwerke.
- **DIN DKE SPEC 99004:2025-05 – Specification of Operational Design Domain in Rail**: Festlegung von Anforderungen zur Beschreibung und Modellierung der Betriebsumgebung, direkt basierend auf den im Projekt entwickelten ODD-Methoden.

Die OVGU brachte sich sowohl fachlich in die Ausarbeitung der Dokumente ein als auch strategisch in deren Platzierung im Normungsprozess. Durch die aktive Teilnahme an Arbeitsgruppen und Abstimmungen mit Partnern trug sie dazu bei, die Projektergebnisse in öffentlich zugänglicher, anwendungsorientierter Form zu dokumentieren und einen nachhaltigen Beitrag zur Standardisierung sicherheitsgerichteter KI im Schienenverkehr zu leisten.

Vorgegebenes Ziel	Erreichtes Ergebnis
Initiierung und Umsetzung von Standardisierungsaktivitäten	Aktive Mitwirkung an nationalen Normungsaktivitäten im Bahnumfeld
Überführung zentraler Projektergebnisse in normative Dokumente	Inhaltliche Beiträge zu methodischen Grundlagen, Begriffen und Anforderungen
Fokusthemen: Sicherheitsmetriken, ODD, Absicherung KI-basierter Systeme	Fachliche Ausarbeitung in relevanten Abschnitten der DIN SPECs
Beteiligung an konkreten Standardisierungsdokumenten	Mitarbeit an DIN DKE SPEC 99002:2025 und DIN DKE SPEC 99004:2025-05

Vorgegebenes Ziel	Erreichtes Ergebnis
Strategische Platzierung im Normungsprozess	Aktive Teilnahme an Arbeitsgruppen und Abstimmungen mit Partnern
Nachhaltige Verankerung der Projektergebnisse	Veröffentlichung öffentlich zugänglicher DIN SPECs für zukünftige Regelwerke

1.5.e) UAP 5.6 Ergebnisverbreitung

In UAP 5.6 war die OVGU maßgeblich an der Verbreitung der im Projekt *safe.trAIn* entwickelten wissenschaftlichen Ergebnisse beteiligt. Als akademischer Partner lag der Fokus insbesondere auf Publikationen zu zentralen Projektthemen wie OOD Detection, Unsicherheitsquantifizierung, Sicherheitsmetriken und erklärbarer KI.

Mehrere Arbeiten wurden auf renommierten Konferenzen und Workshops wie CVPR, ICML, WACV, SAIAD, SafeComp und WAISE veröffentlicht und teilweise mit Preisen ausgezeichnet. Darüber hinaus stellte die OVGU projektbezogene Open-Source-Komponenten (z. B. Bibliotheken zur OOD Detection und Metrikberechnung) bereit, was die externe Nutzbarkeit der Ergebnisse erhöhte.

Neben der wissenschaftlichen Veröffentlichung beteiligte sich die OVGU an der Kommunikation gegenüber einem breiteren Publikum – etwa durch Beiträge für die Projektwebseite, grafisches Material und begleitende Erklärungen zu den Standardisierungsarbeiten (DIN SPECs). Diese Aktivitäten erfolgten im Rahmen der abgestimmten Kommunikationsstrategie.

Damit trug die OVGU entscheidend zur Sichtbarkeit und nachhaltigen Verankerung der Projektergebnisse in Wissenschaft und Öffentlichkeit bei.

Vorgegebenes Ziel	Erreichtes Ergebnis
Wissenschaftliche Dissemination der Projektergebnisse	Zahlreiche Publikationen zu OOD Detection, Unsicherheitsmetriken, erklärbarer KI
Veröffentlichung auf nationalen und internationalen Konferenzen	Beiträge auf CVPR, ICML, WACV, SAIAD, SafeComp, WAISE (teilweise ausgezeichnet)
Beitrag zur Verwertbarkeit der Ergebnisse	Open-Source-Komponenten für OOD Detection und Metrikberechnung veröffentlicht
Breite Kommunikation außerhalb der Fachcommunity	Beiträge für Projektwebseite, visuelles Material, Erklärungen zu DIN SPECs
Unterstützung der Kommunikationsstrategie des Projekts	Aktive Mitwirkung an abgestimmten Disseminations- und Öffentlichkeitsaktivitäten
Sichtbarmachung der Innovationsleistung	Stärkung der Position von *safe.trAIn* in Wissenschaft und Praxis autonomer Mobilität

1.6 Querschnittsthemen und Sonderaufgaben

Neben den expliziten APs war die OVGU federführend in mehreren Special-Interest-Groups (SIGs) tätig:

- SIG Betriebsumgebung/ODD
- SIG Out-of-Distribution Detection
- SIG Erkennung großer Objekte

Diese SIGs adressierten wichtige, AP-übergreifende Herausforderungen und fungierten als methodische Querschnittsplattformen. Durch ihre Leitung und Mitwirkung konnte die OVGU dazu beitragen, dass methodische Konsistenz über die verschiedenen Projektbereiche hinweg gewahrt blieb.

Auch administrative Aufgaben wie das Verfassen des SOTA-Reports zur Verifikation, die Organisation und Ausrichtung von Konsortialtreffen, sowie die kontinuierliche Koordination der beteiligten Partnergruppen fielen in den Verantwortungsbereich der OVGU.

2 Darstellung der Wichtigsten Positionen des zahlenmäßigen Nachweises

2.1 Personalkosten

Die Personalkosten bildeten den größten Ausgabeposten und spiegeln den projektbedingten wissenschaftlichen Aufwand der OVGU wider. Aufgrund tariflicher Vorgaben und interner Regelungen besitzt die Universität hierbei keinen wesentlichen Spielraum in der Mittelverwendung. Die eingesetzten wissenschaftlichen Mitarbeitenden trugen entscheidend zur Bearbeitung der technisch anspruchsvollen Arbeitspakete bei.

2.2 Reisekosten

Die im Projekt angefallenen Reisekosten waren notwendig, um den im Konsortialvertrag vereinbarten Verpflichtungen zur Dissemination der Forschungsergebnisse sowie zur Konsortialkoordination nachzukommen. Insbesondere die Teilnahme an nationalen und internationalen Fachkonferenzen war erforderlich, um die entwickelten Methoden einem fachkundigen Publikum vorzustellen und damit die Sichtbarkeit und Verwertbarkeit der Ergebnisse zu sichern. Darüber hinaus dienten Dienstreisen zu Konsortialtreffen und Projektworkshops der inhaltlichen Abstimmung mit den Projektpartnern.

2.3 Sonstige Mittel

2.3.a) Hilfswissenschaftler

Zur Unterstützung der Projektumsetzung wurden studentische Hilfskräfte (HiWis) eingesetzt. Ihre Mitarbeit war insbesondere für die Erhebung, Annotation und Auswertung von Testdaten, die Unterstützung bei Implementierungsarbeiten sowie die Vorbereitung von Publikationen unerlässlich. Die Tätigkeit der HiWis stellte eine wichtige operative Ergänzung zur wissenschaftlichen Arbeit der Projektmitarbeitenden dar und war für die fristgerechte Erfüllung der Projektziele notwendig.

2.3.b) Hardware

Für die Durchführung der im Projekt geplanten und umgesetzten Experimente im Bereich des maschinellen Lernens – insbesondere im Zusammenhang mit der Entwicklung und Evaluation von Deep-Learning-basierten Sicherheitsmechanismen – war die Anschaffung leistungsfähiger Hardwarekomponenten (insbesondere Grafikkarten) erforderlich. Diese waren für das Training, die Testung und die Validierung der entwickelten KI-Modelle unverzichtbar und bildeten somit eine technische Grundvoraussetzung für die erfolgreiche Bearbeitung der entsprechenden Arbeitspakete

3 Darstellung des voraussichtlichen Nutzens, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Die im Projekt safe.trAIn erzielten Ergebnisse zeigen bereits während der Projektlaufzeit eine hohe wissenschaftliche und anwendungsbezogene Verwertbarkeit. Im Sinne des fortgeschriebenen Verwertungsplans lässt sich der voraussichtliche Nutzen wie folgt darstellen:

3.1 Wissenschaftliche Verwertung

Ein wesentlicher Bestandteil der Verwertung erfolgte durch die Einbindung der Projektergebnisse in die akademische Lehre. Insbesondere wurden zentrale Inhalte und Methoden aus safe.trAIn in das von der OVGU durchgeführte Seminar „Hot Topics for ML Safety“ integriert, das zweimal im Verlauf der Projektlaufzeit mit Studierenden aus verschiedenen Masterstudiengängen durchgeführt wurde. Die dort vermittelten Inhalte basierten zum Teil direkt auf den im Projekt entwickelten Konzepten, wie z. B. OOD Detection, Unsicherheitsmetriken und Sicherheitsnachweisverfahren für KI-Systeme.

Darüber hinaus wurden mehrere Abschlussarbeiten (Bachelor und Master) mit direktem Bezug zu den Projektthemen betreut. Diese Arbeiten griffe*n* konkrete technische Fragestellungen aus den Arbeitspaketen auf, etwa zur Metrikbewertung, Szenarienmodellierung oder Erklärbarkeit von KI-Systemen. Sie trugen nicht nur zur

Vertiefung einzelner Projektaspekte bei, sondern dienten auch der Nachwuchsförderung im Bereich KI-Sicherheit.

Zudem sind derzeit mehrere Promotionsvorhaben mit Bezug zu safe.trAIn in Bearbeitung. Diese werden die im Projekt begonnenen Forschungsfragen weiterentwickeln und vertiefen, insbesondere im Bereich sicherer KI-Anwendungen in sicherheitskritischen Kontexten.

3.2 Transfer und Weiterverwertung

Die Übertragbarkeit der entwickelten Methoden wird über den Bahnbereich hinaus aktiv geprüft. Insbesondere im Bereich Robotik bestehen vielversprechende Anschlussmöglichkeiten, da auch dort KI-gestützte Perzeptionssysteme unter sicherheitskritischen Bedingungen operieren. Erste interne Machbarkeitsanalysen zur Adaption der entwickelten OOD Detection Methoden für Industrielle Robotersysteme wurden angestoßen.

3.3 Publikationen und Projektkooperationen

Im Rahmen von safe.trAIn entstanden mehrere wissenschaftliche Publikationen, die nationale und internationale Sichtbarkeit erzielten. Dabei wurde teilweise in enger Kooperation mit anderen Projektpartnern publiziert, wodurch nicht nur die wissenschaftliche Qualität, sondern auch die Konsistenz der Ergebnisse im interdisziplinären Kontext gestärkt wurde. Die Veröffentlichungen dokumentieren zentrale Fortschritte des Projekts, etwa in der Laufzeiterkennung von Anomalien, der Verifikation von KI-Funktionen oder der Konzeption sicherheitsgerichteter Architekturen.

Insgesamt zeigt sich bereits zum jetzigen Zeitpunkt ein hoher potentieller Transfernutzen, sowohl in die akademische Ausbildung, die Forschung als auch potenziell in industrielle Folgeanwendungen. Die Projektergebnisse bilden eine tragfähige Grundlage für weitere Forschungsvorhaben sowie für zukünftige Entwicklungen im Bereich vertrauenswürdiger KI.

4 Während der Durchführung des Vorhabens dem ZE bekannt gewordenen Fortschritts auf dem Gebiet des Vorhabens bei anderen Stellen

Während der Durchführung des Vorhabens wurde dem Zuwendungsempfänger ein bedeutender technologischer Fortschritt im Bereich großer KI-Modelle bekannt, insbesondere im Hinblick auf Large Language Models (LLMs) und Vision-Language Models (VLMs). Diese Entwicklungen, die parallel zur Projektlaufzeit an Dynamik gewonnen haben, wurden von der OVGU kontinuierlich beobachtet, bewertet und – wo möglich – in die Projektarbeit integriert.

Ein konkreter Beitrag in diesem Kontext ist die im Rahmen des Projekts entstandene wissenschaftliche Publikation: „*Language Models as Reasoners for Out-of-Distribution Detection*."
Diese Arbeit untersucht den Einsatz von LLMs zur Out-of-Distribution Detection, einem der zentralen technischen Schwerpunkte von safe.trAIn. Dabei wird gezeigt, dass LLMs in der Lage sind, Weltwissen und allgemeinsprachlich formulierte Domänenregeln zu nutzen, um Eingaben außerhalb des Trainingsverteilungsbereichs zu erkennen. Die Ergebnisse zeigen, dass LLMs ohne explizite formale Wissensbasen besser als Zufall abschneiden und mit formal basierten Verfahren vergleichbare Leistungen erzielen können, wenn sie durch sprachlich formulierte Domänenkonzepte unterstützt werden. Dieser Ansatz bietet ein gewissen Potenzial für die vereinfachte und flexiblere Gestaltung erklärbarer Sicherheitssysteme.

Auch im Bereich der Vision-Language Models (VLMs) wurde deren Potenzial zur Verbesserung multimodaler Wahrnehmung im Bahnumfeld diskutiert. Allerdings zeigt sich in der praktischen Umsetzung, dass viele aktuelle VLMs zwar leistungsstark und komfortabel in der Anwendung sind, jedoch sehr große Modellgrößen aufweisen. Dies bringt insbesondere im Kontext sicherheitskritischer, ressourcenbeschränkter Echtzeitsysteme erhebliche Herausforderungen mit sich – etwa in Bezug auf Laufzeit, Hardwareanforderungen und Prüfbarkeit. Eine direkte Einbindung solcher Modelle in sicherheitsgerichtete Systeme bleibt daher derzeit nur eingeschränkt möglich und bedarf zusätzlicher Forschung.

Wichtig ist zudem festzuhalten, dass die im Projekt safe.trAIn adressierten Kernprobleme – wie z. B. die strukturierte Sicherheitsnachweisführung, die formale Beschreibung der Betriebsumgebung (ODD), sowie die Entwicklung quantitativer Metriken zur Vertrauenswürdigkeit von KI – nicht durch externe Entwicklungen vollständig gelöst wurden. Trotz des technologischen Fortschritts bestehen weiterhin methodische und regulatorische Lücken, die die sichere Integration moderner KI-Modelle in kritische Infrastrukturen erschweren.

Die im Projekt entwickelten Konzepte und Methoden bleiben daher hochrelevant und zukunftsfähig – nicht zuletzt als robuste Grundlage für die systematische Einbettung neuer KI-Technologien wie LLMs und VLMs in vertrauenswürdige Architekturen.

Im Rahmen des Projekts safe.trAIn wurden zahlreiche wissenschaftliche Arbeiten sowie zwei Standardisierungsdokumente veröffentlicht, die die erarbeiteten Ergebnisse dokumentieren, verbreiten und zur weiteren Verwertung in Wissenschaft und Praxis bereitstellen. Nachfolgend wird eine Übersicht der wichtigsten Veröffentlichungen gegeben, jeweils mit einer kurzen inhaltlichen Einordnung in den Projektkontext:

5 Erfolgte oder geplante Veröffentlichungen des Ergebnisses

5.1 Wissenschaftliche Veröffentlichungen

Continuous Development and Safety Assurance Pipeline for ML-Based Systems in the Railway Domain

SafeComp Workshop

Diese Arbeit beschreibt die Umsetzung eines SafeMLOps-Prozesses im Bahnkontext, wie er im Projekt safe.trAIn entwickelt wurde. Sie zeigt auf, wie ein Git-zentrierter Entwicklungsworkflow zur kontinuierlichen Validierung und Absicherung von KI-Funktionen beiträgt. Die Konzepte wurden direkt im Projektkontext implementiert und mit realen Komponenten getestet.

Improving Out-of-Distribution Detection with Markov Logic Networks

ICML

Diese Publikation verbindet klassische OOD-Detection mit probabilistischer logischer Inferenz durch Markov Logic Networks. Die Arbeit geht direkt aus den Aktivitäten im Bereich Laufzeitüberwachung und erklärbarer Sicherheitsmetriken hervor und zeigt, wie regelbasierte und datengetriebene Systeme kombiniert werden können, um die Verlässlichkeit zu erhöhen.

Out-of-Distribution Detection with Adversarial Outlier Exposure

CVPR Workshop

Hier wird ein Verfahren zur OOD-Erkennung vorgestellt, bei dem synthetische Outlier durch adversariales Training gezielt zur Verbesserung der Robustheit genutzt werden. Die Methode wurde im Projektkontext für sicherheitskritische Bildklassifikation untersucht und trägt zur Absicherung der Wahrnehmungskomponenten bei.

Language Models as Reasoners for Out-of-Distribution Detection

SafeComp Workshop

Diese Arbeit untersucht, wie LLMs zur Laufzeit OOD-Fälle anhand natürlichsprachlicher Regeln erkennen können. Sie stellt eine Brücke zwischen sicherheitsgerichteter

Wahrnehmung und dem aktuellen Stand der Sprachmodell-Forschung dar und zeigt auf, wie KI-Weltwissen zur Verbesserung des Laufzeit-Monitorings genutzt werden kann.

Out-of-Distribution Detection with Logical Reasoning
WACV

Diese Publikation präsentiert einen hybriden OOD-Ansatz, bei dem neuronale Wahrnehmung mit symbolischem Wissen kombiniert wird. Das System bietet neben höherer Leistung auch verbesserte Erklärbarkeit, was für sicherheitskritische KI-Systeme von großer Bedeutung ist.

Towards Deep Anomaly Detection with Structured Knowledge Representations
SafeComp Workshop

Diese Arbeit untersucht, wie strukturierte Wissensrepräsentationen genutzt werden können, um Anomaliedetektion robuster, transparenter und erklärbarer zu machen. Sie stellt eine methodische Grundlage für spätere Entwicklungen im Projekt dar.

Evaluating and Increasing Segmentation Robustness in CARLA
SafeComp Workshop

In dieser Arbeit wurde die Robustheit von Segmentierungsmodellen in simulierten Umgebungen analysiert – ein wichtiges Thema für die sichere KI-Perzeption im Bahnverkehr. Die Erkenntnisse flossen u.a. in die Testszenarien und Evaluierungen im virtuellen Testfeld ein.

5.2 Standardisierungsdokumente

DIN DKE SPEC 99002 – *Terminology – AI in Railway Applications*
Diese Spezifikation definiert zentrale Begriffe für den Einsatz von KI im Bahnumfeld und wurde in enger Abstimmung mit den Projektpartnern – u.a. unter Mitwirkung der OVGU – erstellt. Sie schafft die Grundlage für konsistente Kommunikation und zukünftige Regelwerke.

DIN DKE SPEC 99004 – *Specification of Operational Design Domain in Rail*
Basierend auf der im Projekt entwickelten Methodik zur Beschreibung von Betriebsumgebungen (ODD), formuliert dieses Dokument ein anwendungsnahes Framework zur ODD-Spezifikation im Bahnkontext. Die OVGU wirkte an der Entwicklung und Abstimmung mit.

Die Vielzahl an Veröffentlichungen zeigt die hohe wissenschaftliche Produktivität der OVGU im Rahmen von safe.trAIn. Die Arbeiten decken zentrale Themen wie OOD Detection, Sicherheitsmetriken, Laufzeitüberwachung, erklärbare KI und robuste Wahrnehmung ab und leisten einen nachhaltigen Beitrag zur Verwertbarkeit, Normung und Weiterentwicklung sicherer KI-Systeme im Bahn- und Mobilitätsbereich.