SURE-AD

Trustworthy Scenario Understanding: Robust and Explainable AI for Safe Autonomous Driving

Programm / Ausschreibung	FORPA, Dissertaionen 2024, Industrienahe Dissertationen 2026	**Status**	laufend
Projektstart	01.05.2026	**Projektende**	30.04.2029
Zeitraum	2026 - 2029	**Projektlaufzeit**	36 Monate
Projektförderung	€ 109.928		
Keywords	Trustworthy AI, Autonomous Vehicles, Autonomous Driving, Scenario Understanding		

Projektbeschreibung

Moderne Mobilität ist in starker Transformation und autonome Fahrzeuge zählen zu den wichtigsten Treibern. Eine breite Einführung setzt jedoch voraus, dass autonome Fahrzeuge im realen Betrieb sowohl leistungsfähig als auch sicher sind und erklärbar agieren. Künstliche Intelligenz (KI) ermöglicht in diesem Zusammenhang Situationsverständnis auch in komplexen urbanen Situationen und bildet so eine zentrale Schnittstelle zwischen Wahrnehmung und Entscheidungsfindung innerhalb der Fahrzeugregelung. Aktuelle Ansätze fokussieren dabei auf Vorhersagegenauigkeit, während die Transparenz der Entscheidungsprozesse oft noch unzureichend ist. Insbesondere im Kontext sicherheitskritischer Systeme, wie autonomer Fahrzeuge, die auf transparente und überprüfbare Entscheidungsprozesse angewiesen sind, muss diese Herausforderung zukünftig besser adressiert werden. Das entsprechende Themenfeld „Explainable AI“ (XAI) gewinnt daher in Forschung, Industrie und Regulierung zunehmend an Bedeutung.

Im praktischen Einsatz sind KI-Modelle für das Situationsverständnis in autonomen Fahrzeugen mit Herausforderungen konfrontiert. Die interne Entscheidungslogik dieser Modelle ist nur begrenzt zugänglich, was die Interpretierbarkeit ihrer Vorhersagen erheblich einschränkt und zu mangelnder Transparenz führt. Diese Intransparenz verstärkt strukturelle Schwächen: Seltene, aber sicherheitsrelevante Ereignisse sind in Trainingsdaten häufig unterrepräsentiert, was die Zuverlässigkeit in solchen Grenzfällen reduziert. Gleichzeitig müssen KI-Modelle unvollständige und verrauschte Sensordaten verarbeiten und die Dynamik sowie Multimodalität realer Verkehrsszenarien erfassen, die sich nicht vollständig durch Karten oder Verkehrsregeln abbilden lassen. Bestehende Ansätze, die sich auf Vorhersagegenauigkeit konzentrieren und dabei Aspekte wie Robustheit und Erklärbarkeit vernachlässigen, führen in weiterer Folge zu Herausforderungen in der Validierung und verhindern eine Zertifizierung.

Das Projekt adressiert diese Herausforderungen durch die Entwicklung robuster und erklärbarer Modelle für das Situationsverständnis in automatisierten Fahrzeugen. Das Projekt verfolgt vier Ziele: Erstens werden Modellarchitekturen entworfen, die intrinsische Erklärbarkeit mit hoher Vorhersagegenauigkeit verbinden. Zweitens werden Bewertungsprotokolle entwickelt, die Robustheit und Erklärbarkeit systematisch erfassen. Drittens werden Post-hoc-Methoden für mehr Transparenz so weiterentwickelt und operationalisiert, dass sie für unterschiedliche Stakeholder verständliche und relevante Einblicke liefern. Viertens werden Modelle, Bewertungsmethoden und Erklärungsansätze in ein

integriertes, sicherheitsorientiertes Validierungskonzept überführt, das Transparenz erhöht und Validierungsprozesse effizienter gestaltet. Die Innovation des Projekts liegt in der Verbesserung der Transparenz entlang des gesamten Entwicklungs- und Testzyklus und in der gezielten Nutzung des Situationsverständnisses als zentrale Grundlage für erklärbare und nachvollziehbare Entscheidungen und Handlungen autonomer Fahrzeuge.
Insgesamt trägt das Projekt dazu bei, die Lücke zwischen Leistungsfähigkeit und Erklärbarkeit in autonomen Systemen zu schließen, vor allem durch den gezielten Einsatz von XAI-Methoden. Es stärkt vertrauenswürdige KI in der Mobilität, indem es technische Leistungsfähigkeit mit regulatorischen und gesellschaftlichen Anforderungen vereint und damit sichere und transparente autonome Fahrzeuge ermöglicht.

Abstract

The rapid development of autonomous vehicles (AVs) is transforming mobility, yet widespread adoption depends on robustness and transparency in real-world operation. Current AV systems employ artificial intelligence (AI) for scenario understanding, a core component linking perception, decision-making, and control. However, many approaches still prioritise predictive accuracy over explainability, making it difficult to verify predictions in safety-critical applications. This limitation is especially acute in automated driving (AD), where regulators, industry, researchers, and society require AI systems that are transparent and suitable for rigorous safety assurance.
Scenario understanding models face major challenges in real-world deployment. A central issue is the limited insight into the model's internal decision-making, which restricts the interpretation and justification of predictions. This lack of transparency also complicates solutions to broader challenges. Safety-critical events, such as rare interactions or unexpected behaviours, are underrepresented in datasets, leaving models vulnerable in edge cases. At the same time, models must cope with incomplete or noisy perception data and account for the complex, multi-modal nature of traffic, where even accurate maps and rules cannot fully constrain all possible outcomes. These factors complicate validation and certification, as current evaluation protocols often emphasise accuracy while overlooking robustness and actionable explanations. By improving model interpretability, this project aims to support more trustworthy AD across diverse scenarios.
Motivated by these challenges, the project aims to develop robust, explainable models for scenario understanding in AD. It addresses four central goals: (i) designing model architectures that enhance explainability without sacrificing predictive performance; (ii) establishing evaluation protocols that extend beyond standard accuracy; (iii) operationalising post-hoc explanation methods to provide actionable stakeholder insights; and (iv) integrating models, evaluation protocols, and explanation views into an assurance-oriented evidence package that improves traceability and reduces validation effort. The project's novelty lies in redefining safe scenario understanding for AD by embedding transparency throughout model development and testing. By positioning scenario understanding as a central explanation interface in the AV stack, this research lays the foundation for integrating explainability at the heart of AD and enabling system-wide insight.
As AV technology matures, consensus is growing among regulators, industry, and researchers that trustworthy AD must provide not only stronger performance but also deeper model insights. This project advances trustworthy AI in mobility by prioritising societal safety and meeting evolving requirements, paving the way for safer and more transparent AD.

Projektpartner

- Virtual Vehicle Research GmbH