Methodik zur Absicherung von KI im Fahrzeug

Der Einsatz von KI ist ein Schlüsselelement auf dem Weg zum automatisierten Fahren. Initiiert durch den VDA erarbeitet ein Konsortium aus OEMs, Zulieferern, Technologieprovidern und Forschungseinrichtungen eine Methodik für eine neuartige Sicherheitsargumentation, die systematisch Schwächen von KI-basierten Funktionen identifiziert, messbar macht und mitigiert. Mit dem Projekt „KI Absicherung“, das aus der VDA-Leitinitiative hervorgeht, soll ein Industriekonsens für ein methodisches Vorgehen geschaffen werden, das am Beispiel der Fußgängererkennung aufgezeigt wird.

AUTOREN

Dipl.-Ing. Frédérik Blank ist Senior Project Manager im Bereich Deep Learning Loop for Automated Driving bei der Robert Bosch GmbH in Abstatt.

Dr.-Ing. Fabian Hüger ist Wissenschaftler bei der Volkswagen AG und Lead Expert AI Safety bei der Cariad SE in Wolfsburg.

PD Dr. Michael Mock ist Senior Data Scientist am Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) in Sankt Augustin.

Dr. rer. nat. Thomas Stauner ist Senior Engineer bei der BMW Group in München.

© Mackevision Medien Design | Qualityminds

1 GESAMTANSATZ ZUR ABSICHERUNG

Der Nachweis der funktionalen Sicherheit bei Modulen, bei denen Algorithmen auf Basis künstlicher Intelligenz (KI) zum Einsatz kommen, ist im internationalen Wettbewerb von entscheidender Bedeutung. Deshalb wird im deutschen Projekt „KI Absicherung“ [1], das von der Forschungsvereinigung Automobiltechnik (FAT) als Teil des Verbands der Automobilindustrie (VDA) begleitet wird, eine Methodik erarbeitet, mit der inhärente Schwachstellen von KI-Funktionen systematisch identifiziert und mitigiert werden können. Das Ziel ist die Ableitung einer stringenten evidenzbasierten Sicherheitsargumentation. „KI Absicherung“ ist Teil der Verbundprojekte der KI-Familie, die in dieser ATZ-Ausgabe mit einem Artikel über die VDA-Leitinitiative Autonomes und vernetztes Fahren vorgestellt wird.

BILD 1 zeigt die Spezifikation und Entwicklungsschritte einer KI-Funktion sowie die Methodik zum Aufbau einer evidenzbasierten Sicherheitsargumentation. Die Methodik stützt sich auf Sicherheitsmaßnahmen, Metriken und Tests, die bei der Entwicklung und Validierung zur Anwendung kommen. Die Spezifikation der KI-Funktion ist der elementare Ausgangspunkt, sowohl für die Entwicklung der Funktion selbst als auch für den Aufbau der Sicherheitsargumentation. Neben den rein funktionalen Anforderungen wie der Erkennung von Personen auf Kamerabildern wird der Einsatzbereich, die sogenannte Operational Design Domain (ODD), abgegrenzt. Durch die ODD-Spezifikation kann eine systematische und repräsentative Auswahl von Trainings- und Testdaten für das maschinelle Lernen (ML) mit Deep Neural Networks (DNNs) getroffen werden. Für die detaillierte Spezifikation von Daten und Metadaten werden Beschreibungssprachen und eine Ontologie entwickelt. Sie sind sowohl für Menschen verständlich, um eine nachvollziehbare Sicherheitsargumentation aufbauen zu können, als auch maschinenlesbar, um Datenanalysen und Testauswertungen automatisiert durchführen zu können.

DNNs können als komplexe Blackbox-Approximationsfunktionen verstanden werden, die durch Trainingsdaten optimiert werden. Als solche können sie Schwachstellen in der Generalisierungsfähigkeit aufweisen, die im ungünstigen Fall zu einer Unzulänglichkeit der Softwarefunktion führen. Zur systematischen Adressierung wurde eine Liste von DNN-spezifischen Sicherheitsbedenken erarbeitet, **BILD 2**. Sie beschreiben mögliche Ursachen für funktionale Unzulänglichkeiten, die bei der Sicherheitsbetrachtung zu fokussieren sind. Der Nachweis der hinreichenden Mitigation der DNN-spezifischen Sicherheitsbedenken erfolgt in der Sicherheitsargumentation durch Zusammenbringen aller Evidenzen. Der daraus resultierende Sicherheitsnachweis (Assurance Case) wird unter Verwendung der Goal Structuring Notation (GSN) dokumentiert.

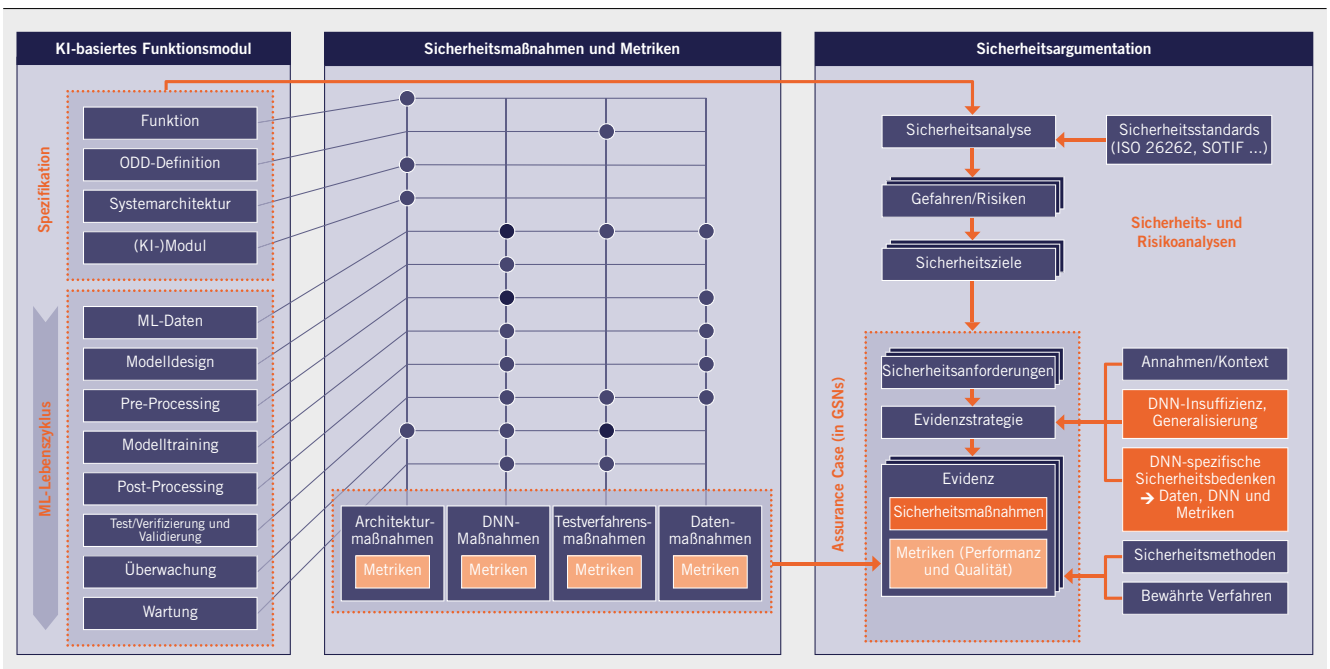

BILD 1 Gesamtansatz zur Absicherung von KI-Funktionen [2] (© BMW | Bosch | Fraunhofer IAIS | Volkswagen)

DNN-spezifische Sicherheitsbedenken		
1 Unzureichende Generalisierungsfähigkeit	1.4 Mangelnde Plausibilität	2.4 Unzureichende Spezifikation der ODD
1.1 Unzuverlässige Konfidenzschätzung	2.1 Die Datenverteilung ist keine gute Annäherung an die reale Welt.	2.5 Veränderungen des Eingaberaums über die Zeit
1.2 Mangelnde Robustheit	2.2 Unzureichende Trennung von Test- und Trainingsdaten	2.6 Unbekanntes Verhalten in seltenen kritischen Situationen
1.2.1 Mangelnde zeitliche Stabilität	2.3 Abhängigkeit von der Labelingqualität	3.1 Unzureichende Berücksichtigung von Sicherheitsaspekten in Performanzmetriken
1.3 Unverständliches Verhalten	2.3.1 Fehlende Labelingdetails oder Meta-Labels	--- Mangelnde (algorithmische) Effizienz

Funktionale Unzulänglichkeit | Bedenken bezogen auf DNN-Merkmale | Bedenken bezogen auf Daten | Bedenken bezogen auf Metriken | Andere

BILD 2 DNN-spezifische Sicherheitsbedenken [3] (© BMW | Bosch | Continental | Fraunhofer IAIS | Volkswagen)

2 ONTOLOGIEBASIERTE SPEZIFIKATION

Neben der Funktionsspezifikation bildet die Definition der ODD eine entscheidende Voraussetzung für die Herleitung eines Sicherheitsnachweises. Eine ausreichende Generalisierungsfähigkeit und Performanz von DNNs ist nur dann erreichbar, wenn es keine essenziellen Lücken in den Trainingsdaten gibt, und nachweisbar, wenn die Testdaten die ODD repräsentativ abdecken. Grundsätzlich kann die ODD aus der Sicht von Felddaten betrachtet werden sowie komplementär dazu aus der Sicht einer semantischen Strukturierung des Eingaberaums, wobei beide zu berücksichtigen sind. Die Felddatensicht umfasst alle im vorgesehenen Einsatzbereich beobachteten Daten. Die Sicht der semantischen Strukturierung definiert hingegen den Eingaberaum und formalisiert diesen durch eine darauf abgestimmte Ontologie.

Eine Ontologie für die Funktion Fußgängererkennung wurde im Projekt iterativ definiert, **BILD 3**. Sie besteht aus zehn Domänen, die unter anderem Personen- und Objekteigenschaften, Licht-

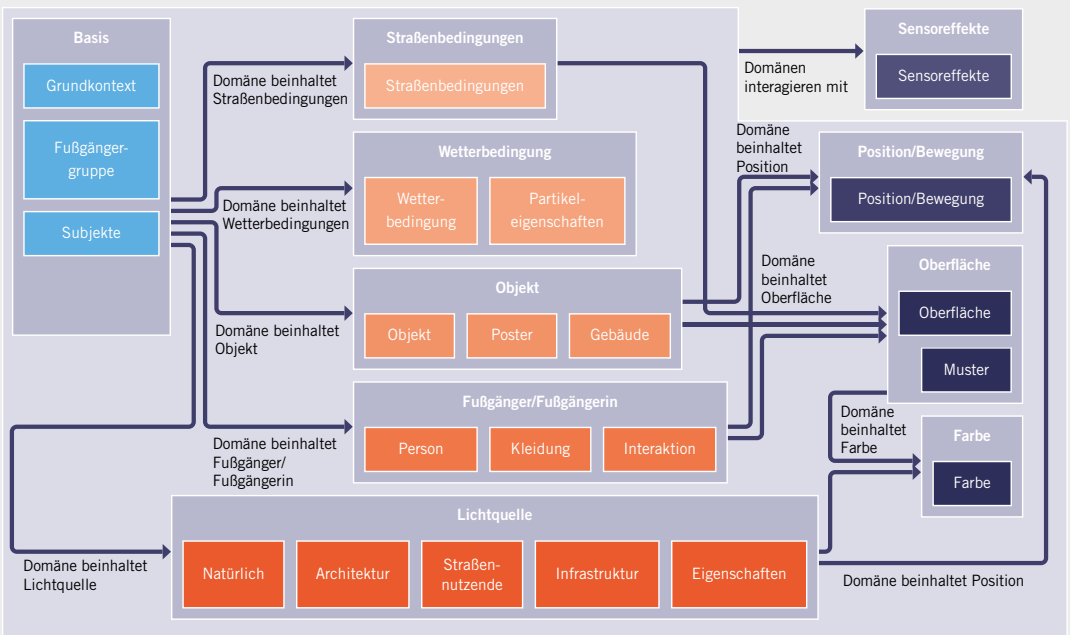

BILD 3 Struktur der Ontologie zur Fußgängererkennung, bestehend aus zehn Domänen [4] (© Bosch)

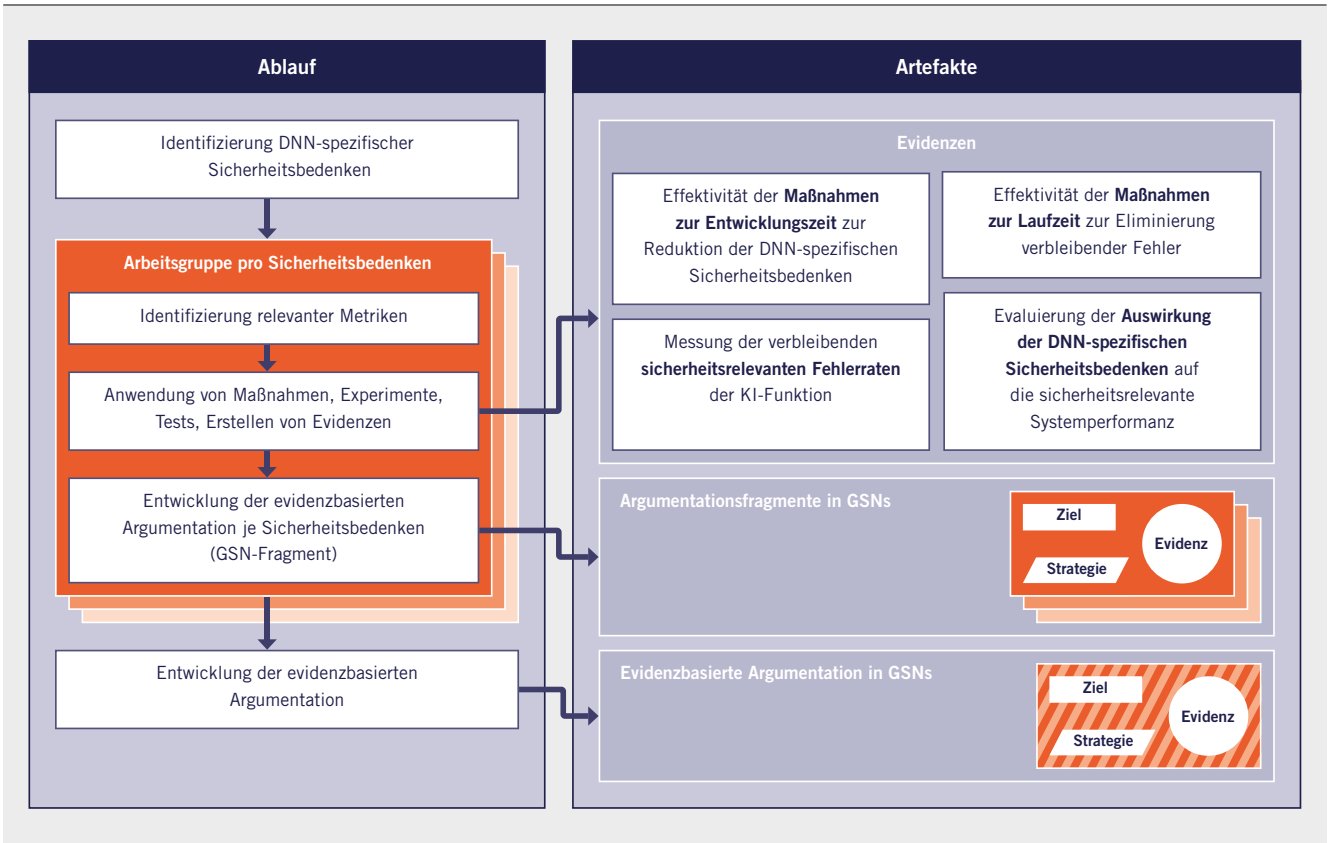

BILD 4 Erstellung einer evidenzbasierten Sicherheitsargumentation (© Bosch | Fraunhofer IKS | Volkswagen)

effekte und Wetter adressieren. Ausgangspunkt war eine Vorstrukturierung des Eingaberaums anhand von initialen Domänen, die in Experteninterviews ausgearbeitet wurden. Basierend auf der Ontologie kann zum einen die ODD als Teil des Eingaberaums formal beschrieben und zum anderen Anforderungen an die Trainings- und Testdaten bezüglich darin vorkommender Ausprägungen oder Kombinationen von Ausprägungen der Subdomänen definiert werden. Die Daten können hinsichtlich der Abdeckung der ODD analysiert werden, wobei deren Repräsentativität nicht durchgehend die gleiche Granularität aufweisen muss. Höhere Granularität ist für diejenigen Domänen nötig, die die Performanz der KI-Funktion wesentlich beeinflussen, wie zum Beispiel der Verdeckungsgrad von Fußgängern oder Fußgängerinnen. Relevante Einflussfaktoren werden im Laufe eines Entwicklungsprojekts iterativ verfeinert, indem Datenbereiche mit schlechter DNN-Performanz tiefer untersucht werden. Solche semantisch orientierten DNN-Performanzanalysen zeigen dabei nicht nur physikalisch bedingte Einflussfaktoren auf, sondern auch Schwächen, die durch fehlende Daten im Trainingsprozess entstehen.

3 METHODIK

Den Sicherheitsstandards ISO 26262 und ISO 21448 folgend ist für die Freigabe einer hochautomatisierten Fahrfunktion ein Sicherheitsnachweis erforderlich, der in Form einer Sicherheitsargumentation mit systematisch generierten Evidenzen aufzeigt, dass Unzulänglichkeiten der KI-Funktion hinreichend mitigiert (worden) sind. Da die Fehlerraten einer KI-Funktion im „Open-World"-Kontext nicht vollständig gemessen werden können, wird bei der Methodik auf ein Framework mit einer systematischen evidenzbasierten Sicherheitsargumentation gesetzt, **BILD 4**. Analog zum Kausalmodell der ISO 21448 werden bekannte Unzulänglichkeiten mit geeigneten Maßnahmen adressiert [5]. Das Erreichen der Sicherheitsziele und -anforderungen wird argumentiert anhand der Effektivität dieser Maßnahmen, der verbleibenden Fehlerraten und des Einflusses der Unzulänglichkeiten auf die sicherheitsrelevante Systemperformanz.

Das Framework zur Mitigation DNN-spezifischer Sicherheitsbedenken unterscheidet zwischen Maßnahmen zur Entwicklungszeit (zum Beispiel bei der Datenerzeugung/-auswahl und Training beziehungsweise Test) und Maßnahmen zur Laufzeit. Während der Entwicklungszeit wird beispielsweise die Robustheit der Funktion gegenüber Veränderungen im Eingangssignal (wie Farbveränderungen oder Unschärfen) gemessen und wenn nötig mit geeigneten Verfahren verbessert. Die Messungen selbst können als Evidenzen für die Sicherheitsargumentation herangezogen werden – ebenso die durch geeignete Maßnahmen erzielte Steigerung der Robustheit. Zur Laufzeit werden unter anderem Überwacher zur Erkennung von bisher unbekannten, sogenannten Out-of-Distribution-Eingangsdaten und/oder Unsicherheiten verwendet, um die zur Entwicklungszeit unbekannten Einflussfaktoren mit zu berücksichtigen. Die Funktionalität der Überwacher wird getestet und die Ergebnisse als Evidenzen für die Effektivität der Maßnahmen verwendet.

Zur Messung der sicherheitsrelevanten Fehlerraten werden spezifische Metriken verwendet, die Sicherheitsaspekte wie zum

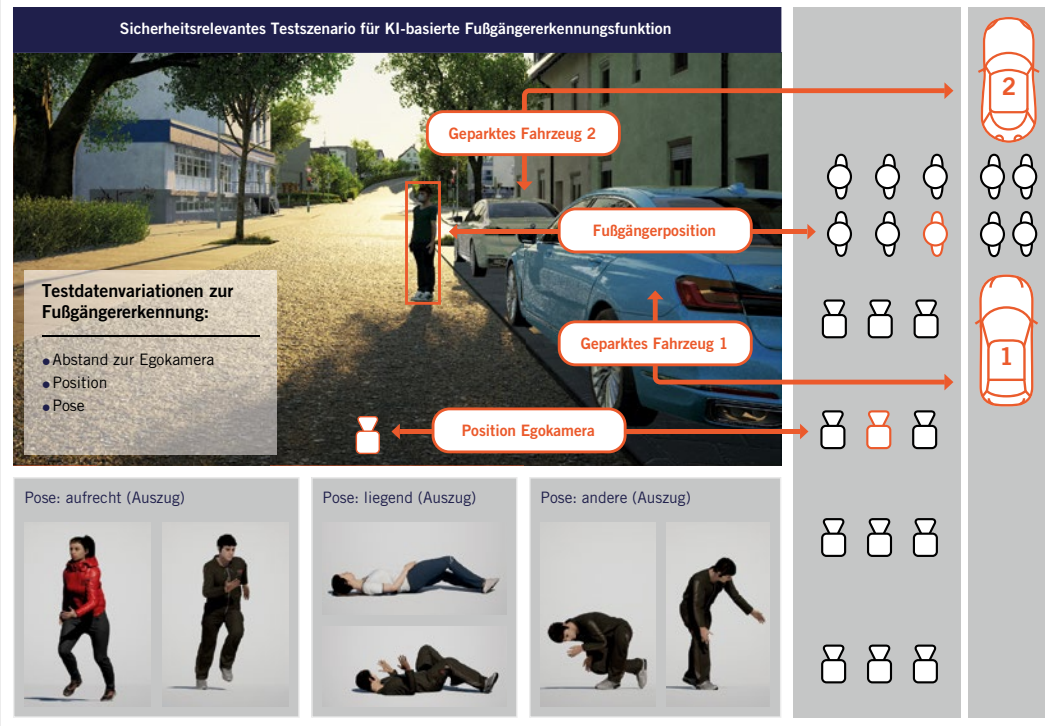

BILD 5 Parametrisierbares sicherheitsrelevantes Testszenario der Fußgängererkennung (© Bosch | Mackevision Medien Design)

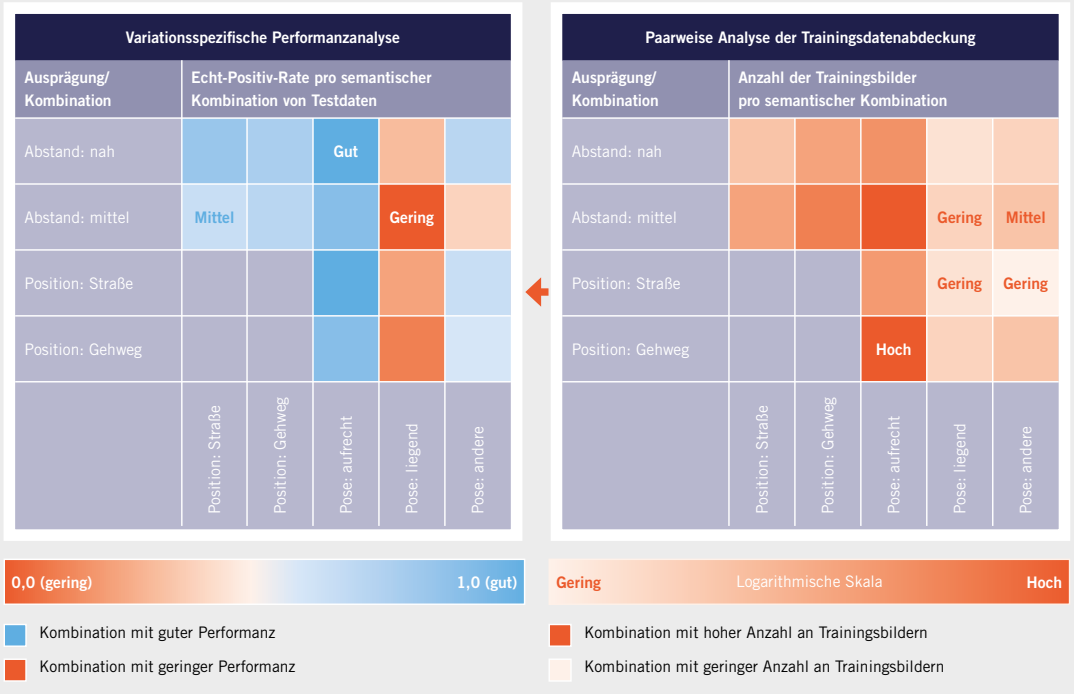

BILD 6 Variationsspezifische Performanzanalyse auf der Grundlage von Daten aus dem Testszenario (links) und paarweise Analyse der Trainingsdatenabdeckung (rechts) (© Bosch)

Beispiel die Entfernung eines Fußgängers oder einer Fußgängerin zum Egofahrzeug und seine beziehungsweise ihre Position innerhalb oder außerhalb eines angenommenen Bremswegs und relativ zur Straße berücksichtigen. So können sicherheitsrelevante Akzeptanzkriterien sowie entsprechende Tests formuliert werden.

Als Entwicklungsansatz in diesem Projekt zur Erzeugung der Maßnahmen und Evidenzen wird in sogenannten Evidenz-Workstreams gearbeitet. Dabei wird für verschiedene DNN-spezifische Sicherheitsbedenken jeweils eine Arbeitsgruppe aus KI-Entwickelnden und KI-Testenden sowie Absicherungsfachleuten gebildet. In

jeder Gruppe werden jeweils geeignete Metriken und Tests definiert sowie priorisierte Maßnahmen angewendet und ausgewertet. Parallel zur Funktion werden Fragmente der Sicherheitsargumentation entwickelt. Sind die mitigierenden Maßnahmen hinreichend wirksam, so werden diese in die Sicherheitsargumentation integriert.

4 ANWENDUNGSBEISPIEL

Die unzureichende Abdeckung der ODD durch geeignete Trainings- beziehungsweise Testdaten entspricht dem DNN-spezifischen Sicherheitsbedenken unter Punkt 2.1 aus **BILD 2**. Ontologiebasierte Analysen in Verbindung mit kombinatorischem Testen [6] liefern einen wichtigen Baustein zur Bestimmung der systematischen Eingaberaumabdeckung. Dieses Verfahren ist nutzbar, um Lücken in Datensätzen zu identifizieren beziehungsweise gezielt daraus Anforderungen an neue Testdaten abzuleiten.

Zur Veranschaulichung zeigt **BILD 5** den Auszug eines im Projekt entwickelten Testszenarios, das hier lediglich die drei Parameter Fußgängerabstand, -position und -pose beinhaltet. Diese werden systematisch variiert und miteinander kombiniert. Die Nutzung von synthetischen Daten (im Projekt) erlaubt die Erstellung von systematisch parametrierten Testszenarien auf Grundlage der definierten Ontologie. Bei Anwendung des hier im Beispiel definierten und kombinatorisch ausgeglichenen Testdatensatzes auf das zu untersuchende DNN zeigen sich Performanzlücken bei bestimmten sicherheitsrelevanten Parameterkombinationen, **BILD 6** (links). Dies ist für die Pose der liegenden Person im mittleren Abstand der Fall. Mittels einer Analyse der Trainingsdatenabdeckung, **BILD 6** (rechts), konnten fehlende Posendaten als eine relevante Ursache für den Performanzabfall identifiziert und die DNN-Performanz in diesem Bereich durch ein DNN-Retraining mittels Anreicherung der Trainingsbilder mit Posendaten von liegenden Fußgängern beziehungsweise Fußgängerinnen sichtbar verbessert werden.

Die Methode der Datenabdeckungsanalyse sollte sowohl mit weiteren Untersuchungen (wie Sensitivitäts- und Robustheitsanalysen) als auch dezidierten Testverfahren für eine Nutzung als Evidenz in der Sicherheitsargumentation kombiniert werden. Die Robustheit der KI-Funktion kann durch Tests mit systematisch veränderten Eingangsdaten, durch veränderte Sensorabbildungen oder aber durch systematisches Aufbringen von Bildstörungen (wie Rauschen oder Verzerrungen) argumentiert werden. Auf der Trainingsseite können weitere relevante Daten auch durch zusätzliche Maßnahmen wie beispielsweise Active Learning identifiziert werden. Zur Identifikation von Unzulänglichkeiten zur Laufzeit können zusätzlich Maßnahmen wie zum Beispiel Unsicherheitsbewertungen eingesetzt werden.

5 ZUSAMMENFASSUNG

Eine der größten Herausforderungen bei der Integration von KI-basierten Algorithmen in hochautomatisiert fahrende Fahrzeuge ist, die Sicherheit der Funktionsmodule nachweisen zu können. Die im Projekt erarbeitete neue Methodik beschreibt ein ganzheitliches und iteratives Vorgehen für eine evidenzbasierte Sicherheitsargumentation – bezogen auf die Sicherheit einer beabsichtigten Funktionalität (Safety Of The Intended Functionality, SOTIF). Zur Generierung nutzbarer Evidenzen wurden Absicherungs- und Testmethoden und Maßnahmen entwickelt und integriert. Die Projektergebnisse werden unter anderem in die Kommunikation mit Standardisierungsgremien wie ASAM und ISO/PAS 8800 eingebracht.

LITERATURHINWEISE

[1] European Center for Information and Communication Technologies (EICT) GmbH (Hrsg.): KI Absicherung – Safe AI for Automated Driving. Online: https://www.ki-absicherung-projekt.de, aufgerufen: 23. Februar 2022

[2] Mock, M. et al.: An Integrated Approach to a Safety Argumentation for AI-Based Perception Functions in Automated Driving. Safecomp: International Conference on Computer Safety, Reliability, and Security, York, September 2021

[3] European Center for Information and Communication Technologies (EICT) GmbH (Hrsg.): Newsletter Nr. 2, KI Absicherung: DNN-specific Safety Concerns. Online: https://ki-absicherung-projekt.de/safety-concerns, aufgerufen: 23. Februar 2022

[4] Herrmann, M. et. al.: Using ontologies for data set engineering in automotive AI applications. DATE, Design, Automation and Test in Europe, online, 2022

[5] Houben, S. et. al.: Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety. Online: https://arxiv.org/pdf/2104.14235.pdf, aufgerufen: 6. April 2022

[6] Gladisch, C. et al.: Leveraging combinatorial testing for safety-critical computer vision datasets. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, Juni 2020

DANKE

Das Forschungsvorhaben „KI Absicherung“ als Teil des Verbundprojekts der KI-Familie, das aus der VDA-Leitinitiative Autonomes und vernetztes Fahren entstanden ist, wurde vom Bundesministerium für Wirtschaft und Klimaschutz (BMWK) aufgrund eines Beschlusses des Deutschen Bundestages gefördert. Es wurde vom Verband der Automobilindustrie (VDA) initiiert und von der Forschungsvereinigung Automobiltechnik e. V. (FAT) begleitet. Für die finanzielle Förderung und die Zusammenarbeit mit Partnern aus Industrie und Forschung sei an dieser Stelle gedankt. Weiterer Dank gilt den Co-Autoren Stephan Scholz (Volkswagen AG), Andreas Rohatschek und Martin Herrmann (Robert Bosch GmbH) sowie Simon Burton (Fraunhofer IKS).