Assurance Methodology for In-vehicle AI

The application of AI is a key enabler for highly automated driving. Initiated by VDA, a consortium of OEMs, suppliers, technology providers and scientific institutions is developing a methodology for a novel safety argumentation in the project "KI Absicherung" (safe AI) that systematically identifies insufficiencies of AI-based functions, makes them measurable and mitigates them. The project stems from the "VDA-Leitinitiative" (flagship initiative). An industrial consensus for a methodical approach is to be achieved which is demonstrated using the example of pedestrian detection.

AUTHORS

Dipl.-Ing. Frédérik Blank is Senior Project Manager Deep Learning Loop for Automated Driving at Robert Bosch GmbH in Abstatt (Germany).

Dr.-Ing. Fabian Hüger is Researcher at Volkswagen AG and Lead Expert AI Safety at Cariad SE in Wolfsburg (Germany).

PD Dr. Michael Mock is Senior Data Scientist at Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in Sankt Augustin (Germany).

Dr. rer. nat. Thomas Stauner is Senior Engineer at BMW Group in Munich (Germany).

© Mackevision Medien Design | Qualityminds

1 OVERALL APPROACH FOR ASSURANCE

Assuring the safety of functions that make use of AI-based algorithms is crucial for the German automotive industry in international competition. Therefore "KI Absicherung" [1], supported by the Research Association for Automotive Technology (FAT) as part of the German Association of the Automotive Industry (VDA), is developing a methodology to systematically identify and mitigate inherent insufficiencies of AI functions. The goal is to derive a stringent evidence-based safety argumentation. This project is part of the collaborative projects of the AI family ("KI Familie"), which is presented in this issue of ATZworldwide with an article about the VDA flagship initiative autonomous and connected driving.

FIGURE 1 shows the specification and development steps of an AI function, as well as the methodology for building an evidence-based safety argumentation. The method is based on safety measures, metrics and tests that are applied during development and validation. The specification of the AI function is the elementary starting point, both for the development of the function itself and for that of the safety argumentation. In addition to the purely functional requirements, such as recognition of persons on camera images, the scope of use of the function, the so-called Operational Design Domain (ODD), is defined. Through the ODD specification, a systematic and representative selection of training and testing data for Machine Learning (ML) with Deep Neural Networks (DNNs) can be made. Description languages and an ontology are developed for the detailed specification of data and metadata. These are understandable for humans to be able to build up a comprehensible safety argumentation, as well as machine-readable to be able to carry out data analyses and test evaluations automatically.

DNNs can be understood as complex black box approximation functions that are optimized by training data. As such, they may have insufficiencies in the generalization capability, which in the unfavorable case can lead to insufficiencies of the software function. To address this systematically, a list of DNN-specific safety concerns has been developed, **FIGURE 2**. They are indications of functional insufficiencies that must be focused on. For these, appropriate mitigation measures are needed. Demonstrating that DNN-specific safety concerns are sufficiently addressed is done in the safety argumentation based on evidence. The resulting assurance case is documented using the Goal Structuring Notation (GSN).

2 ONTOLOGY-BASED SPECIFICATION

In addition to the functional specification, the definition of the ODD is a crucial prerequisite for the derivation of a safety argumentation.

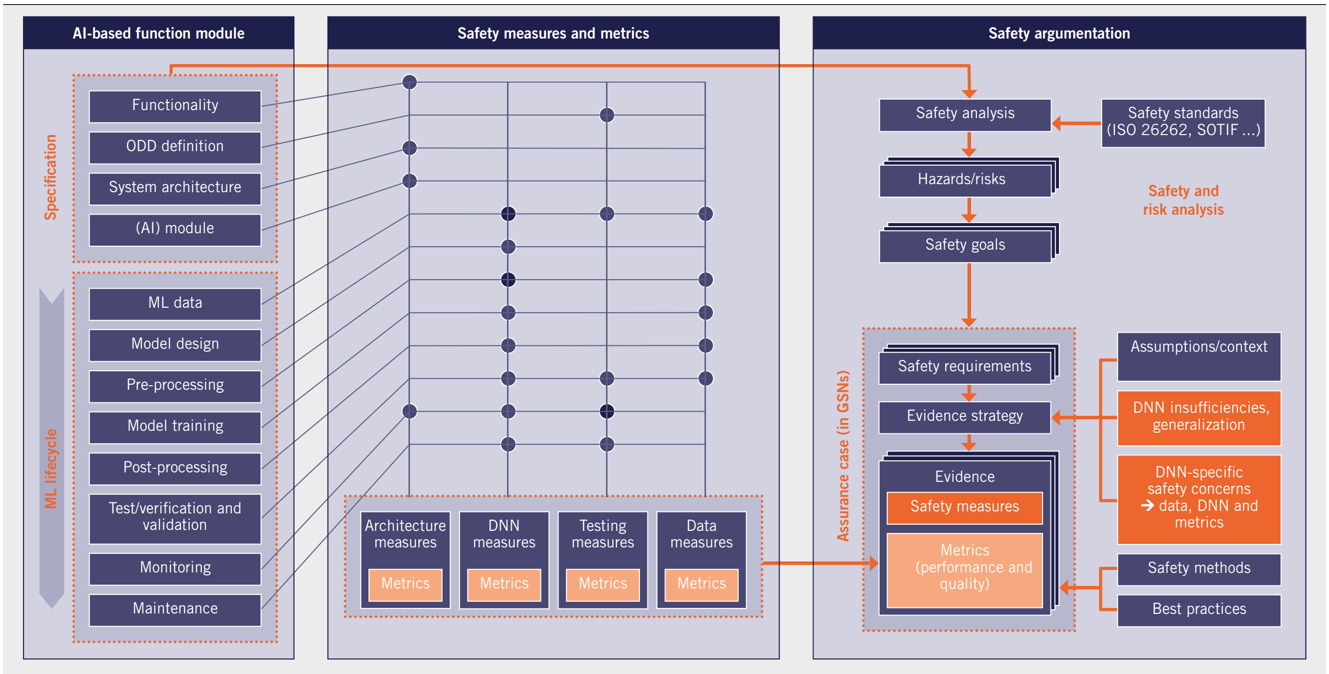
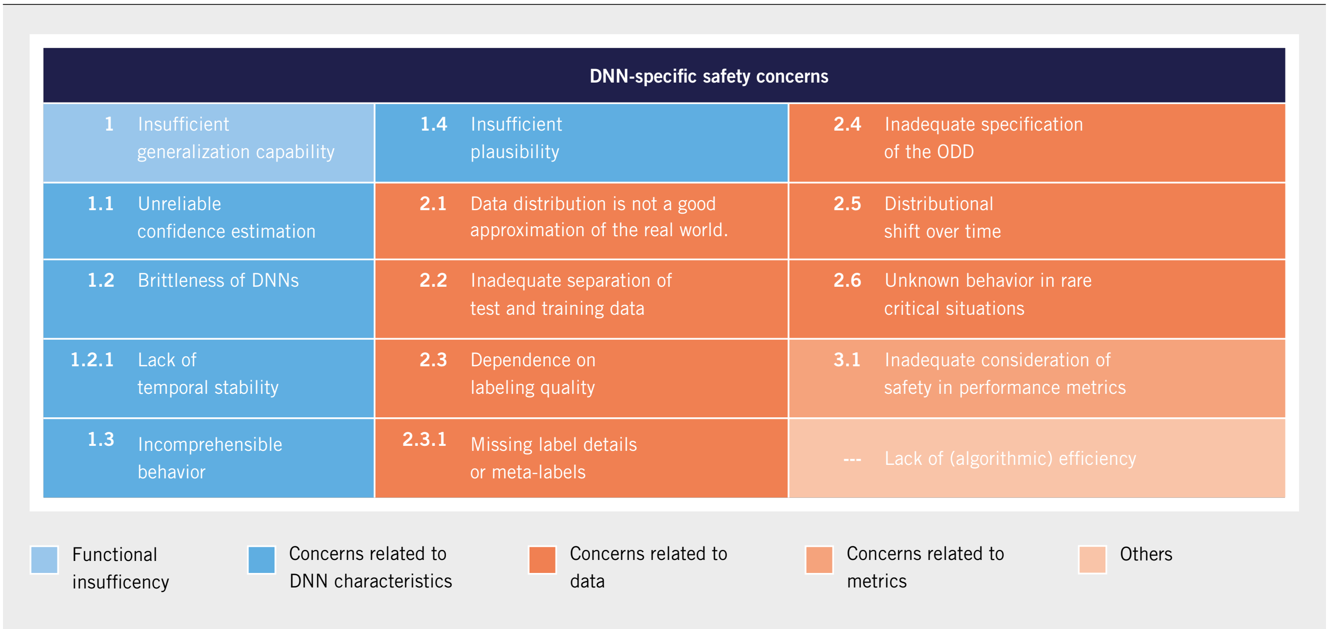


FIGURE 1 Overall approach for assurance of AI-based functions [2] (© BMW | Bosch | Fraunhofer IAIS | Volkswagen)



DNN-specific safety concerns		
1 Insufficient generalization capability	1.4 Insufficient plausibility	2.4 Inadequate specification of the ODD
1.1 Unreliable confidence estimation	2.1 Data distribution is not a good approximation of the real world.	2.5 Distributional shift over time
1.2 Brittleness of DNNs	2.2 Inadequate separation of test and training data	2.6 Unknown behavior in rare critical situations
1.2.1 Lack of temporal stability	2.3 Dependence on labeling quality	3.1 Inadequate consideration of safety in performance metrics
1.3 Incomprehensible behavior	2.3.1 Missing label details or meta-labels	--- Lack of (algorithmic) efficiency

■ Functional insufficency ■ Concerns related to DNN characteristics ■ Concerns related to data ■ Concerns related to metrics ■ Others

FIGURE 2 DNN-specific safety concerns [3] (© BMW | Bosch | Continental | Fraunhofer IAIS | Volkswagen)

Sufficient generalization capability and performance of the DNN are only achievable if there are no essential gaps in the training data, and are only justifiable if the test data covers the ODD representatively. In principle, ODD can be viewed from the perspective of field data as well as complementarily from the perspective of semantic structuring of the input space, taking both into account. The field data view includes all data observed in the intended area of application. The semantic structuring view, on the other hand, defines the input space and formalizes it by means of an ontology that is tailored to it.

An ontology for the pedestrian detection function was iteratively defined, **FIGURE 3**, consisting of ten domains that address, among other things, person and object properties, light effects, and weather. The starting point was a pre-structuring of the input space using initial domains developed in expert interviews. Based on the ontology, the ODD can be formally described as a part of the input space and requirements for the training and test data can be defined with respect to the variations or combinations of variations of the subdomains. The data can be analyzed with regard to the coverage of the ODD, whereas their representativeness does not need to have

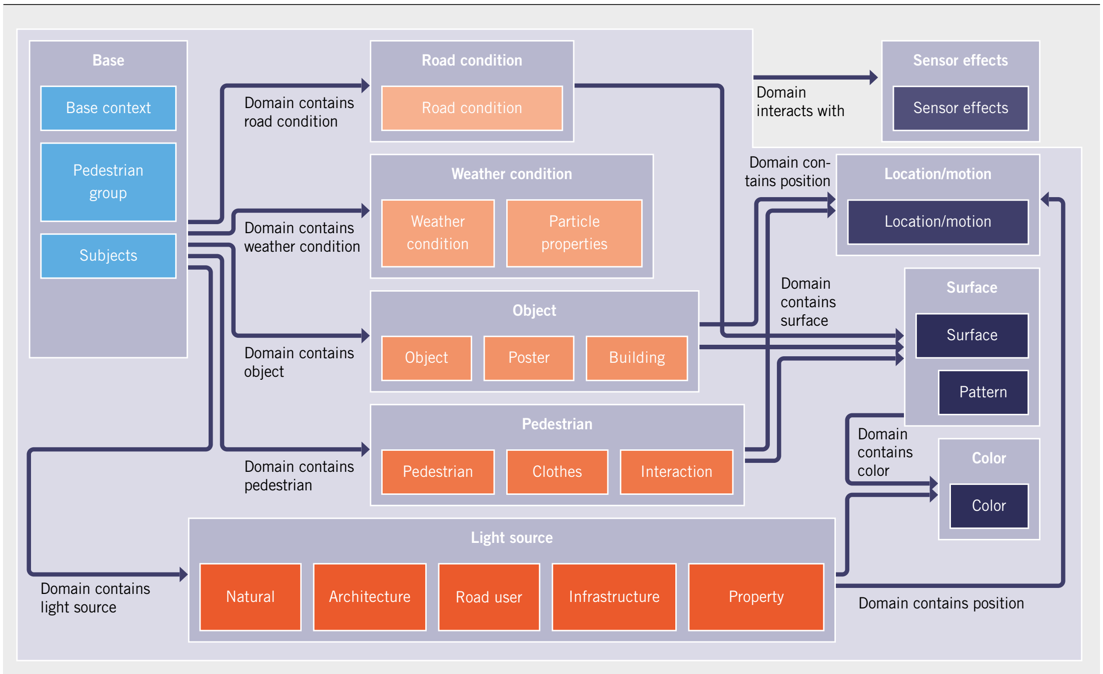


FIGURE 3 Structure of pedestrian detection ontology with ten domains [4] (© Bosch)

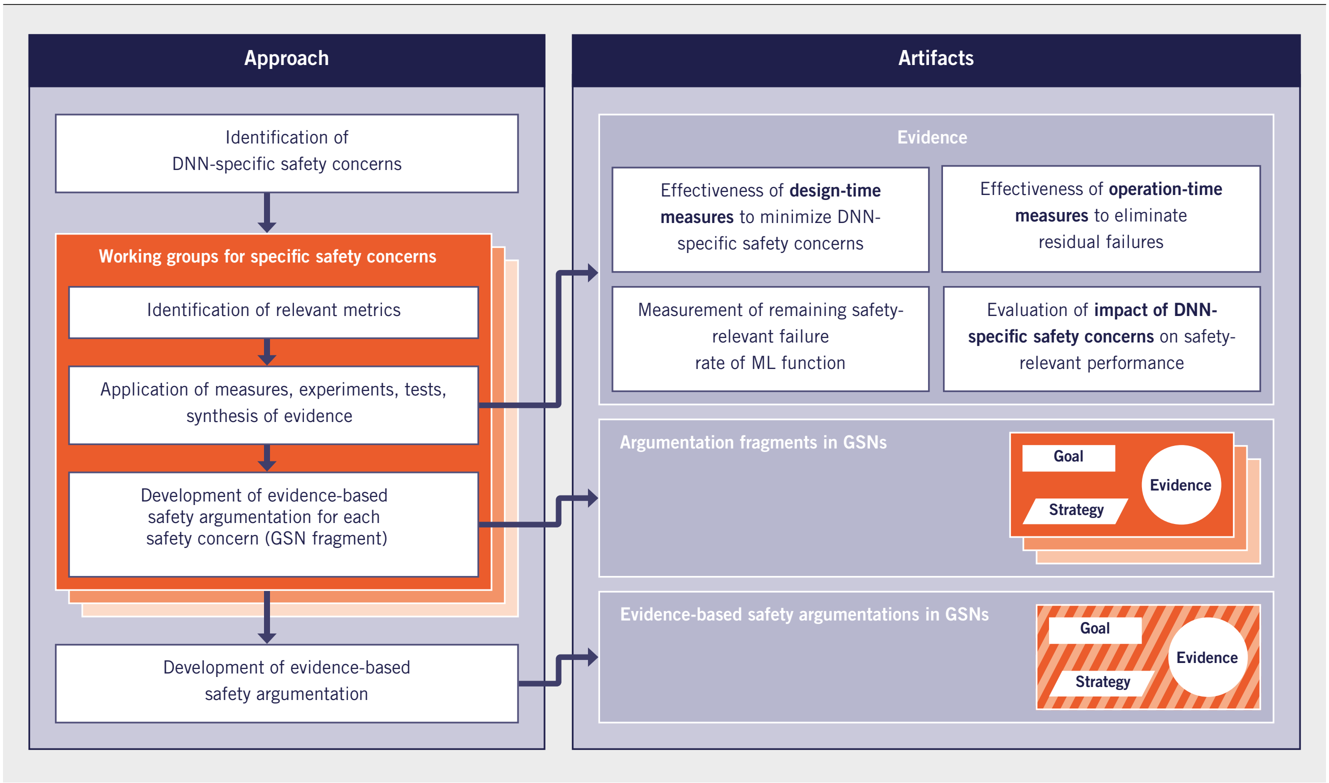

Approach
Artifacts
Identification of DNN-specific safety concerns
Working groups for specific safety concerns
Identification of relevant metrics
Application of measures, experiments, tests, synthesis of evidence
Development of evidence-based safety argumentation for each safety concern (GSN fragment)
Development of evidence-based safety argumentation
Evidence
Effectiveness of design-time measures to minimize DNN-specific safety concerns
Effectiveness of operation-time measures to eliminate residual failures
Measurement of remaining safety-relevant failure rate of ML function
Evaluation of impact of DNN-specific safety concerns on safety-relevant performance
Argumentation fragments in GSNs
Goal
Evidence
Strategy
Evidence-based safety argumentations in GSNs
Goal
Evidence
Strategy

FIGURE 4 Creation of an evidence-based safety argumentation (© Bosch I Fraunhofer IKS I Volkswagen)

the same granularity throughout. Higher granularity is needed for those domains that significantly influence the performance of the AI function, such as the degree of pedestrian occlusion. Relevant influencing factors are refined iteratively during development by in-depth examination of data areas with poor performance. Such semantic-oriented performance analyses do not only reveal physical influencing factors, but also weaknesses that arise from a lack of data in the training process.

3 METHODOLOGY

Following the safety standards ISO 26262 and ISO 21448, the release of a highly automated driving function requires a safety case that demonstrates, in the form of a safety argumentation with systematically generated evidence, that insufficiencies of the AI function have been sufficiently mitigated. Since the error rates of an AI function cannot be fully measured in the “open-world” context, the methodology relies on a framework with a systematic evidence-based safety argumentation, **FIGURE 4**. Analogous to the causal model of ISO 21448, known insufficiencies are addressed with appropriate measures [5]. Arguing that the safety goals and requirements are met is done on the basis of the effectiveness of these measures, the remaining safety-relevant error rates and the influence of the insufficiencies on the safety-relevant system performance.

The framework for mitigation of DNN-specific safety concerns distinguishes between design-time measures (for example, in data generation and selection, training or testing) and operation-time measures. During design time, for example, the robustness of the function against changes in the input signal (such as color changes or blurring) is measured and, if necessary, improved using suitable methods. The measurements themselves can be used as evidence for the safety argumentation – as well as the increase in robustness achieved by appropriate measures. At operation time, monitors are used, among other things, to detect so-called out-of-distribution input data and/or uncertainties in order to take into account the influencing factors unknown at design time. The functionality of these monitors is tested and the test results are used as evidence for the effectiveness of these measures.

For the measurement of the error rate of the pedestrian detection (function), safety aspects are taken into account in the metric calculation such as distance to the ego vehicle and location inside or outside an assumed braking distance and relative to the road. In this way, achievable and safety-oriented criteria as well as corresponding tests can be formulated.

As a development approach within this project for generating these measures and evidences, so-called evidence work streams are used. A working group with AI developers, testers and safety experts is set up for various DNN-specific safety concerns. Within each group, appropriate metrics and tests are defined and prioritized measures are applied and evaluated. Fragments of the safety argumentation are developed in parallel with the functions. If the mitigating measures are sufficiently effective, they are integrated into the safety argumentation.

4 APPLICATION EXAMPLE

The insufficient coverage of the ODD by suitable training or test data corresponds to the DNN-specific safety concern under 2.1 in **FIGURE 2**. Ontology-based analyses of data coverage in conjunc-

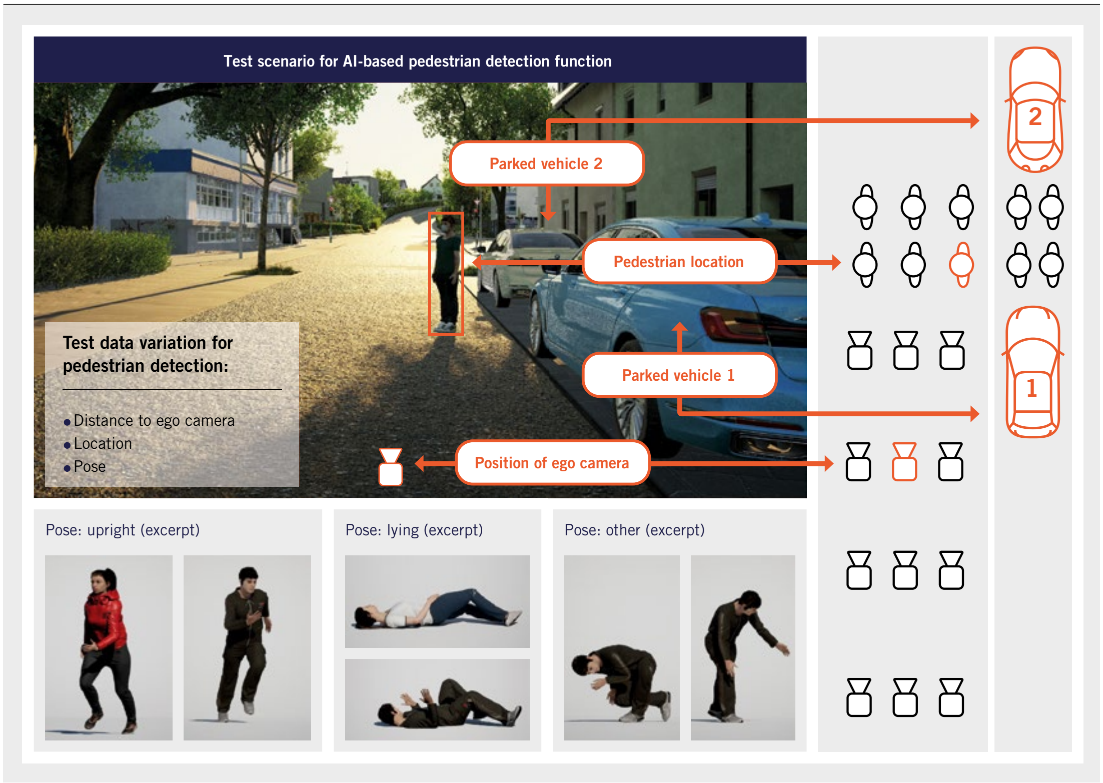
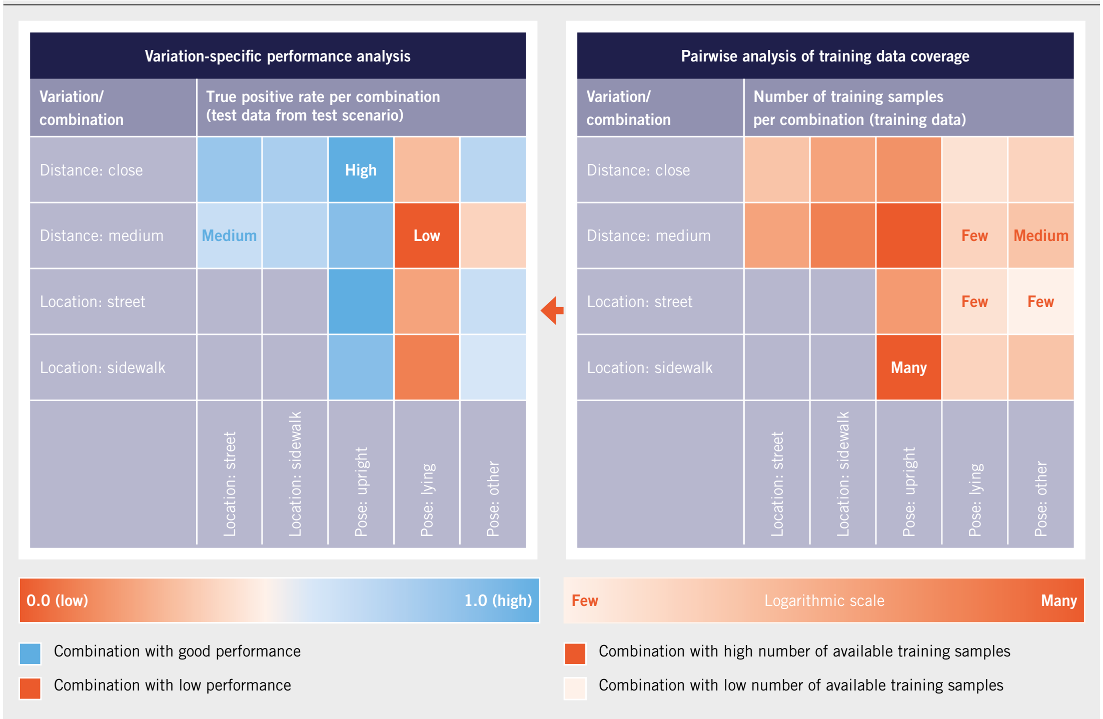


FIGURE 5 Parameterizable safety-related test scenario of pedestrian detection (© Bosch | Mackevision Medien Design)



Variation-specific performance analysis

Variation/ combination	True positive rate per combination (test data from test scenario)				
Distance: close			High		
Distance: medium	Medium			Low	
Location: street					
Location: sidewalk					
	Location: street	Location: sidewalk	Pose: upright	Pose: lying	Pose: other

Pairwise analysis of training data coverage

Variation/ combination	Number of training samples per combination (training data)				
Distance: close					
Distance: medium				Few	Medium
Location: street				Few	Few
Location: sidewalk			Many		
	Location: street	Location: sidewalk	Pose: upright	Pose: lying	Pose: other

🟦 Combination with good performance

🟥 Combination with low performance

🟧 Combination with high number of available training samples

⬜ Combination with low number of available training samples

FIGURE 6 Variations-specific performance analysis based on data from the test scenario (left) and pairwise training data coverage analysis (right) (© Bosch)

tion with combinatorial testing [6] provide an important building block for determining systematic input space coverage. This technique can be used to identify gaps in data sets or to specifically derive new test data requirements from them.

For illustration purposes, **FIGURE 5** shows an excerpt of a test scenario developed in the project, which here only contains the three variable parameters pedestrian distance, location and pose. These are systematically varied and combined. The use of synthetic data (in the project) allows the creation of systematically parameterized test scenarios based on the defined ontology. Applying here in the example the defined and combinatorically balanced test data set to the DNN under investigation, performance gaps for certain safety-relevant combinations are revealed, **FIGURE 6** (left). This is the case for the pose of the person lying in the medium distance. By means of a training data coverage analysis, **FIGURE 6** (right), limited lying pose data could be identified as a relevant cause for the performance drop. The DNN performance in this area was notably improved by a DNN re-training by adding supplementary lying pedestrian image data.

The method of coverage analysis should be combined with further investigations (such as sensitivity and robustness analyses) as well as dedicated test methods for a usage as evidence in the safety argumentation. The robustness of the AI function can be argued by tests with systematically altered input data, altered sensor images or systematic application of image disturbances (such as noise or distortions). On the training side, additional relevant data can also be identified by further measures, for example, such as active learning. Additional measures such as uncertainty quantifications can be used to identify insufficiencies during operation time.

5 SUMMARY

One of the biggest challenges in integrating AI-based algorithms into highly automated vehicles is to be able to assure the safety of the function modules. The new methodology developed in the project describes a holistic and iterative approach of an evidence-based safety argumentation, related to the Safety Of The Intended Functionality (SOTIF). Assurance and testing methods and measures were developed and integrated to generate usable evidence. The project results will also be used in the communication with standardization bodies such as ASAM and ISO/PAS 8800.

REFERENCES

[1] European Center for Information and Communication Technologies (EICT) (ed.): KI Absicherung, Safe AI for Automated Driving. Online: https://www.ki-absicherung-projekt.de, access: February 23, 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) (ed.): Newsletter No. 2, KI Absicherung: DNN-specific Safety Concerns. Online: https://ki-absicherung-projekt.de/safety-concerns, access: February 23, 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, access: April 06, 2022

[6] Gladisch, C.; Heinzemann, C.; Herrmann, M.; Woehrle, M.: Leveraging combinatorial testing for safety-critical computer vision datasets. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, June 2020

THANKS

The research project "KI Absicherung" as part of the joint project of the "KI-Familie," which arose from the VDA flagship initiative autonomous and connected driving, was funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK) based on an enactment of the German Parliament. It was initiated by the German Association of the Automotive Industry (VDA) and supported by the German Research Association for Automotive Technology (FAT). The authors would like thank for the financial support and the cooperation with partners from industry and research. Further thanks go to the co-authors Stephan Scholz (Volkswagen AG), Andreas Rohatschek and Martin Herrmann (Robert Bosch GmbH) and Simon Burton (Fraunhofer IKS).