Defining Required and Feasible Test Coverage for Scenario-Based Validation of Highly Automated Vehicles*

Christian Amersbach, Hermann Winner

Abstract—**A statistical, distance-based validation of highly automated vehicles is not feasible due to the high required testing distance. Scenario-based validation approaches promise to solve this issue. However, due to the high number of influence parameters, the number of possible parameter combinations is exploding. Therefore, exhaustive testing of all possible combinations is not feasible as well. Thus, a coverage criterion for scenario-based validation is required. Hereby, it is crucial that all stakeholders accept the coverage criterion. This paper proposes an approach to determine the number of scenarios that correspond to the required testing distance of the known distance-based approach. Furthermore, the number of scenarios that can be feasibly simulated for validation is estimated under certain assumptions. Comparing the required and the feasible number of scenarios shows that there is still a gap of around one order of magnitude. Nevertheless, combining this approach with other methods that aim to reduce the approval effort has the potential to get the required test coverage to a feasible level and therefore contribute to solving the validation challenge. However, there are still many remaining challenges, such as the availability of representative scenario catalogs or sufficient simulation models for environment perception sensors.**

I. Introduction

One of the major challenges for the release of highly automated vehicles (HAV, Level 3 and higher according to [1]) is their safety approval. First of all, the question "How safe is safe enough?" has not been answered in detail yet. Junietz et al. [2] propose to define macroscopic safety requirements based on risk acceptance of the involved stakeholder groups. They conclude that risk acceptance depends on the market share of HAV and is different for each stakeholder group. Starting from a 10 % market share, the risk acceptance of the society is dominant and requires HAV to be at least 1.3 times safer as today's traffic. With increasing market share, the safety requirement increases as well. Liu et al. [3] state that HAV have to be 4-5 times safer than today's traffic based on a survey study. In any case, it is obvious that HAV have to be safer than today's traffic even if the exact factor is not known yet.

Once a generally accepted safety requirement has been found, the next challenge is to prove that it is fulfilled by a specific HAV in its operational design domain (ODD). Especially if the safety reference is already on a high level of safety, statistical safety approval is not feasible. For example, today's traffic on the German Autobahn is quite safe and the average driven mileage between two fatal accidents is around $7 \cdot 10^8$ km [4, p. 23, 5]. To statistically prove that a HAV, that is assumed to be twice as safe as its reference, is in fact safer than this reference, the reference distance has to be driven ten times according to Wachenfeld and Winner [6].

A scenario-based approach promises to solve this approval challenge. However, it introduces a new challenge, the so-called parameter space explosion: Due to the high number of parameters, the number of possible parameter combinations is exploding. Thus, it is impossible to include all theoretically possible scenarios in the safety validation process. Therefore, it has to be defined and argued which test coverage is sufficient for the scenario-based validation of HAV. This is the focus of this paper.

After introducing the scenario-based approach and related work, the remainder of this paper defines the required number of scenarios that correspond to the distance-based approach. Furthermore, it is estimated how many scenarios can be feasibly tested in simulation. Additionally, measures to close the gap between the required and feasible number of scenarios are proposed and finally the results and their shortcomings are discussed and open research needs are pointed out.

II. Related Work

A. Scenario-based Approach

Schuldt et al. [7] motivate a scenario-based validation approach for HAV, which is also the subject of various research projects [e.g. 8–10]. The main idea behind this approach is to generate *relevant scenarios* intentionally in simulations or on proving grounds rather than waiting until they randomly occur in public traffic. *Relevant scenarios* are all scenarios that can occur within the ODD and are challenging for the HAV and therefore might cause failures [11].

Ulbricht et al. [12] define the term scenario as following:

"A scenario describes the temporal development between several scenes […] Actions & events as well as goals & values may be specified to characterize this temporal development in a scenario. Other than a scene, a scenario spans a certain amount of time."

Bagschik et al. [13] propose a 5-layered model for scene description that is adapted from [14]. Menzel et al. [15] define three levels of detail for the description of scenarios: The most detailed level are *concrete scenarios* that are described by a concrete parameter set and can be transferred to test cases. *Logical scenarios* use parameter ranges for the description

*This Research is part of the project PEGASUS and was founded by the German Federal Ministry for Economic Affairs and Energy (BMWi).

Christian Amersbach and Hermann Winner are with the Institute of Automotive Engineering at Technische Universität Darmstadt, GERMANY (e-mail: amersbach@fzd.tu-darmstadt.de, winner@fzd.tu-darmstadt.de).

while *functional scenarios* use a semantic description in natural language.

Scenarios can be extracted from existing data such as field operational tests (FOT), natural driving studies (NDS) or accident databases by criticality metrics [e.g. 11, 16] or can be artificially generated with different approaches [e.g. 13, 17–20]. A central database to store relevant scenarios is proposed in [21].

B. *Parameter Space Explosion*

As a multitude of N parameters $p_i, i \in \{1 \dots N\}$ is needed to describe a scenario unambiguously and each parameter can take v_i different values, the number of possible combinations within the parameter space is exploding. N hereby depends on the complexity of the scenario and the detail level of its description while v_i depends on the parameter range and its discretization. According to Grindal et al. [22], the resulting size of the parameter space S_N, i.e. the number of all possible parameter combinations, can be calculated as follows:

$$S_N = \prod_{i=1}^{N} v_i \tag{1}$$

Table 1 summarizes the values for N and S_N of some parameter spaces for logical scenarios available in literature. These examples show the broad ranges for N and S_N. Considering the fact that each of the examples represents only one single logical scenario, it becomes obvious that exhaustive testing (ET, i.e. testing all possible scenarios) is not feasible [23, 24]. The parameter space explosion is analyzed in more detail in [25].

C. *Combination Strategies and Random Sampling*

As ET is not feasible for complex systems such as HAV, systematic combination strategies are required to generate test suites. Grindal et al. [22] give an overview of different combination strategies. An alternative to combination strategies are random sampling methods such as Monte-Carlo-Simulation. Arcuri and Briand [24] compare combination strategies and random sampling for the detection of interaction faults, i.e. faults that are caused by a combination of interacting parameters that are typical for complex systems [26]. They conclude that random sampling may outperform combination strategies for high values of N. However, if constraints among the parameters exist, which is the case for the validation of HAV (e.g. there is a constraint between velocity and curvature), combination strategies have an advantage over random sampling as the constraints can be included in the combination strategies.

TABLE I. EXAMPLES FOR THE PARAMETER SPACE EXPLOSION

Reference	Type of Scenario	Number of Parameters N	Number of possible combinations S_N
[27, Annex F]	example scenario for SOTIF safety analysis	8	$2 \cdot 10^{7}$
[28]	test scenario for a lane departure warning (LDW)	16	$3 \cdot 10^{8}$
[25]	exemplary Autobahn scenario “following”	18	$4 \cdot 10^{18}$
[25]	exemplary Autobahn scenario “swiss scenario”	33	$8 \cdot 10^{30}$

Both, random sampling methods and combination strategies, have in common that either the total number of required test cases or a coverage criterion is required as input for the algorithms.

A commonly used criterion for test coverage is t-wise coverage, which is defined by [22] as following:

For t-wise coverage “[…] every possible combination of all […] values of t parameters [has to] be included in at least one test case in the test suite.”

According to [25], the size of a test suite that follows the t-wise criterion is

$$S_t = \prod_{i=1}^{t} \max_i(v_1, \dots, v_N)\,. \tag{2}$$

Here, $\max_i (v_1, \dots, v_N)$ is defined as the i-greatest element of the set $(v_1, \dots, v_N)$.

III. Required Test Coverage

One crucial part of the approval process for HAV is the safety argumentation. In order to reach a general acceptance and confidence, it has to be argued that the performed approval is correct and sufficient to prove that the HAV is safe enough. This argumentation would already be a challenge for ET as the detail level of scenario description (i.e. the number of parameters that describe a scenario) and parameter discretization are more or less arbitrarily chosen. Nevertheless, they have a huge influence on the size of the parameter space. As ET is not feasible, the argumentation becomes even more difficult. For t-wise coverage, the sufficient value for t has to be chosen, which is – in software testing - typically done based on empirical data analysis [26]. However, for HAV there are not enough data available yet to perform empirical analyses. And data of today’s traffic are not directly transferable for HAV as they do not behave similarly to human drivers in all circumstances. For example, we assume HAV would drive more defensive than most human drivers and therefore will stimulate more cut-in maneuvers. The same problem emerges when defining threshold values for criticality metrics, which are not necessarily the same for HAV and human drivers. Additionally, the application of a criticality metric includes the simulation of the scenarios to be examined and is therefore as time-consuming as using the scenarios for simulation-based validation. For other test suite generation methods such as randomly distributed test suites (e.g. Monte Carlo experiments), the required test coverage has to be defined and argued to be valid as well.

As there is – to the best knowledge of the authors – currently no method existing that specifies the required test coverage for scenario-based validation of HAV, the distance-based, statistical approach proposed by Wachenfeld and

Winner [6] as well as Kalra and Paddock [29] is transferred to the scenario-based approach.

As an example for the following considerations, we use an HAV to be driven on the German Autobahn. Therefore, the average distance between two fatal accidents on the Autobahn is used as safety reference, as it is the rarest (and therefore most difficult to prove) and most severe accident type. In 2017, 356 fatal accidents occurred on the Autobahn [4, p. 23] while the total driven distance was 246 billion kilometers [5]. This leads to a reference distance of $\bar{s}_{\mathrm{ref}} \approx 7 \cdot 10^8$ km. In order to transfer this distance to a reference number of concrete scenarios n_{ref}, we need to estimate the average distance that is covered by one scenario $\bar{s}_{\mathrm{sc}}$. Therefore, we assume the average timely duration of one scenario $\bar{\tau}_{\mathrm{sc}}$ as well as the average velocity $\bar{v}_{\mathrm{sc}}$:

$$\bar{s}_{\mathrm{sc}} = \bar{\tau}_{\mathrm{sc}} \cdot \bar{v}_{\mathrm{sc}} \tag{3}$$

In our example we assume $\bar{\tau}_{\mathrm{sc}} \approx 7.5$ s and $\bar{v}_{\mathrm{sc}} \approx 30\ \frac{\mathrm{m}}{\mathrm{s}}$. Thus, we get $\bar{s}_{\mathrm{sc}} \approx 225$ m.

However, as the dividing line between two consecutive scenarios cannot be clearly specified (e.g. the transition between lane change and following is fluent) there is some overlap between two consecutive scenarios. Thus, we introduce an overlap factor f_{o} to calculate the total number of concrete scenarios n_{tot} that is included in $\bar{s}_{\mathrm{ref}}$:

$$n_{\mathrm{tot}} = f_{\mathrm{o}} \frac{\bar{s}_{\mathrm{ref}}}{\bar{s}_{\mathrm{sc}}} \tag{4}$$

As it is nearly impossible to estimate the average time or distance fraction with overlapping scenarios, the conservative assumption that there are always two overlapping scenarios, i.e. $f_{\mathrm{o}} = 2$, is made. Furthermore, the fact that one specific concrete scenario might be included multiple times in $\bar{s}_{\mathrm{ref}}$ has to be considered. Therefore, a uniqueness factor f_{u} is defined:

$$f_{\mathrm{u}} = \frac{n_{\mathrm{u}}}{n_{\mathrm{tot}}} \tag{5}$$

Hereby n_{u} is the number of unique concrete scenarios within n_{tot}. To quantify f_{u}, a clear definition for the uniqueness of scenarios as well as an adequate sample of scenarios would be required. As both are not available yet, the conservative assumption of $f_{\mathrm{u}} = 1$ is used. With these assumptions, n_{ref} is determined:

$$n_{\mathrm{ref}} = f_{\mathrm{o}} f_{\mathrm{u}} \frac{\bar{s}_{\mathrm{ref}}}{\bar{s}_{\mathrm{sc}}} \approx 6 \cdot 10^9 \tag{6}$$

According to [6], $\bar{s}_{\mathrm{ref}}$ has to be tested with a distance factor $f_{\mathrm{d,95}}$ in order to reach a significance level of $e = 5\ \%$. In the case of an HAV that is twice as safe as the reference, this factor is approximately 10. Transferring this to our approach while considering the conservative estimations of f_{o} and f_{u}, the upper limit for the number of concrete scenarios that are required for the validation is

$$n_{\mathrm{req}} = f_{\mathrm{d,95}} n_{\mathrm{ref}} \approx 6 \cdot 10^{10}. \tag{7}$$

Having determined n_{req}, it can be directly used as a target value for the generation of randomly distributed test suites. If a *t*-wise test coverage is preferred, n_{req} can be used to select the required value for *t* so that

$$S_t \geq n_{\mathrm{req}} \tag{8}$$

is fulfilled. For the two last examples from table 1, the application of (8) would require at least *6*-wise coverage, considering that the mentioned examples are not the only functional scenarios within the test suite. The size of the *t*-wise test suite S_t hereby not only depends on *t* but also on the discretization of the parameter space (c.p. (2)). Therefore, *t* cannot be defined independently from the parameter discretization. Thus, the compromise between fine discretization and a high value of *t* has to be solved. A coarser discretization will result in a higher value for *t* and thus failures with a higher FTFI (failure triggering fault interaction) number can be unveiled [26]. However, a too coarse discretization might lead to undiscovered failures in the parameter space between the sampling points.

IV. Feasible Test Coverage

To estimate which test coverage is feasible, some additional assumptions need to be made. First of all, the total time τ_{tot} that can be spent for the validation has to be defined. We assume that one year seems to be feasible for τ_{tot}. As in the previous section, we use $\bar{\tau}_{\mathrm{sc}} \approx 7.5$ s. It is assumed that all tests can be carried out in simulation, thus we need to estimate feasible values for the real-time factor f_{rt} and the parallelization factor f_{p}. The real-time factor strongly depends on the detail level of the simulation and the computational hardware used. Herby, a realistic simulation of the environment perception with active sensors, such as radar and lidar, is the most computationally intensive and leads to low values for f_{rt} around one in the best case [30]. With simplified or ideal sensor models, higher values for f_{rt} can be achieved at the expense of reduced fidelity and validity. Those models are therefore not sufficient for all validation tests [31]. Thus, we assume $f_{\mathrm{rt}} = 1$ for our estimation. The attainable parallelization factor is only limited by the used computational hardware and could, in theory, be increased to infinity. However, this is not feasible in practice. We assume $f_{\mathrm{p}} = 1000$ might be feasible as it lies in the same order of magnitude as the number of prototypes that are used for the validation of current vehicles [32]. With these assumptions, the feasible number of test scenarios can be estimated:

$$n_{\mathrm{f}} = f_{\mathrm{rt}} f_{\mathrm{p}} \frac{\tau_{\mathrm{tot}}}{\bar{\tau}_{\mathrm{sc}}} \approx 4 \cdot 10^9 \tag{9}$$

V. Closing the Gap Between Required and Feasible Test Coverage

When comparing (7) and (9), it becomes obvious that the required number of concrete test scenarios is higher than the feasible one in our example:

$$n_{\mathrm{req}} \approx 15\, n_{\mathrm{f}} \tag{10}$$

Nevertheless, this seems to be a solvable challenge. The trivial but not most feasible solution would be to increase the parallelization factor f_{p} and/or the total validation time τ_{tot} in order to increase the number of feasible scenarios n_{f}. Instead of increasing n_{f}, decreasing n_{req} seems to be more efficient. To do so, the factors f_{o} and f_{u} have to be estimated more realistically rather than assuming conservative values as it was done in this study. Therefore and first of all, a representative set of scenarios has to be available. NDS recorded by drones such as the highD dataset [33] could be empirically analyzed for this purpose. However, despite the fact that the data set contains more mileage and individual vehicles than comparable studies, it is still too small for a statistically significant analysis and as it only contains good weather conditions and a small amount of Autobahn segments it is not representative for today's traffic in general. Furthermore, HAV behave differently compared to human drivers and therefore cause different interactions with other traffic participants. Accordingly, only the data recorded from drivers or ADAS that behave similarly to HAV (e.g. defensive drivers) are representative. Scenarios generated in complex traffic simulation toolchains as proposed by Hallerbach et al. [20] can be an alternative to recorded scenarios. However, as all simulated scenarios, they bear some uncertainties and imperfections and the suitability for the generation of representative scenarios set has not been proven yet. Once a representative set of scenarios is available, a clear definition for the start and endpoint of a concrete scenario is required for an analysis of the typical overlap between consecutive scenarios. As a result, f_{o} can be estimated more realistically. To define a realistic value for f_{u}, a metric for the uniqueness or equality of scenarios has to be defined in addition to analyzing representative data. We propose the following definition:

Two concrete scenarios are equal if their respective parameter combination is situated in the same volume cell of the common parameter space.

As a threshold value, the cell volume of the volume cell has to be defined. This could be done with a sensitivity analysis.

Furthermore, the approach proposed in this paper to define a sufficient test coverage can be combined with approaches that aim to reduce the validation effort. For instance, the functional decomposition approach introduced by the authors in previous publications [25, 34] aims to reduce the approval effort by decomposing the driving function into functional layers (see Fig. 1) that are tested individually.

In addition to the potential to reduce the validation effort by a factor of around 70 (for 6-wise coverage, cp. Section III) as shown in [25], this approach can also increase the real-time factor as the simulation of environment perception is not required for all functional layers. Thus, combining the functional decomposition approach with the test coverage definition proposed here might lower the required number of test cases below the feasible number of test cases and thus overcome the parameter space explosion.

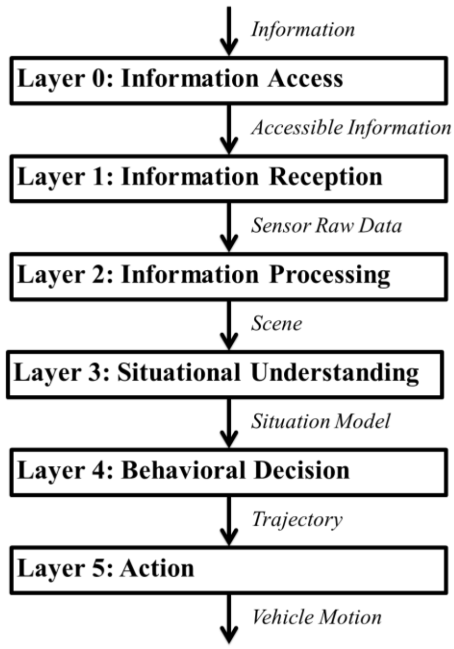

Figure 1. Functional layers and interfaces [25]

VI. Conclusion

In this paper, a method to define the required test coverage for the scenario-based validation of HAV has been proposed and exemplarily applied for the first time. Using the average distance between fatal accidents in the ODD as a safety reference, an existing statistical approach for distance-based validation [6, 29] is transferred to scenario-based validation. With this method, the number of scenarios that are required to create equal evidence as the distance-based approach can be specified. However, to be successfully applicable for the validation of HAV, the method first has to be accepted by all stakeholders [cp. 2]. Furthermore, only the total number of required concrete scenarios has been specified so far. To use this approach for the validation of HAV, the distribution of the total number amongst different functional scenarios (e.g. lane-change, following, approaching a traffic jam, etc.) and within the parameter space has to be specified as well. The probability distribution of today's traffic can only be used conditionally as human drivers behave differently compared to HAV. Complex traffic simulations as proposed in [20] as well as empirical studies that analyze the interaction between human drivers and HAV in the ODD might solve this issue. Additionally, the limits of the validation parameter space have to be defined, as robustness tests require a certain amount of test cases to be outside the specified ODD.

Furthermore, an estimation of the number of scenarios that can be simulated under given assumptions shows that the required number of scenarios seems to be feasible if additional methods to reduce the approval effort are applied. However, this estimation assumes that all test cases can be carried out in simulation, which is actually not the case yet, amongst other reasons due to the lack of accurate models for environment perception sensors [30].

We conclude that this work gives orientation on how to solve the validation challenge for HAV if all stakeholders accept the proposed argumentation. Nevertheless, there are still a lot of unsolved challenges as for example the development of sufficient simulation models or the generation of representative scenario catalogs. However, all of them are currently addressed within various research projects.

References

[1] *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*, SAE J3016, 2014

[2] P. Junietz, U. Steininger, and H. Winner, "Macroscopic Safety Requirements for Highly Automated Driving", *Transportation Research Record*, 2019

[3] P. Liu, R. Yang, and Z. Xu, "How Safe Is Safe Enough for Self-Driving Vehicles?", *Risk analysis*, 2018

[4] Statistisches Bundesamt (Destatis), *Verkehrsunfälle - Zeitreihen 2017*, 2018

[5] Bundesanstalt für Straßenwesen, *Entwicklung der gesamten Fahrleistung von Kraftfahrzeugen auf Autobahnen in Deutschland von 1990 bis 2017*. [Online] Available: https://de.statista.com/statistik/daten/studie/155732/umfrage/fahrleistung-auf-autobahnen-in-deutschland/. Accessed on: Mar. 20 2019

[6] W. Wachenfeld and H. Winner, "The Release of Autonomous Vehicles", in *Autonomous Driving: Technical, Legal and Social Aspects*, H. Winner, M. Maurer, J. C. Gerdes, and B. Lenz, Eds., Berlin, Heidelberg: Springer, 2016, pp. 425–449

[7] F. Schuldt, F. Saust, B. Lichte, M. Maurer, and S. Scholz, "Effiziente systematische Testgenerierung für Fahrerassistenzsysteme in virtuellen Umgebungen", *Automatisierungssysteme, Assistenzsysteme und Eingebettete Systeme Für Transportmittel*, 2013

[8] German Aerospace Center (DLR), *PEGASUS - Project Homepage*. [Online] Available: https://www.pegasusprojekt.de/en/about-PEGASUS. Accessed on: Mar. 04 2019

[9] AVL LIST GMBH, *About the project – Enable S3*. [Online] Available: https://www.enable-s3.eu/about-project/. Accessed on: Mar. 04 2019

[10] D. Zhao *et al.,* "Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling Techniques", (eng), *IEEE transactions on intelligent transportation systems : a publication of the IEEE Intelligent Transportation Systems Council*, vol. 18, no. 3, pp. 595–607, 2017

[11] W. Wachenfeld, P. Junietz, R. Wenzel, and H. Winner, "The worst-time-to-collision metric for situation identification", in *2016 IEEE Intelligent Vehicles Symposium (IV): 19-22 June 2016*, Piscataway, NJ: IEEE, 2016, pp. 729–734

[12] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer, "Defining and substantiating the terms scene, situation, and scenario for automated driving", in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, 2015, pp. 982–988

[13] G. Bagschik, T. Menzel, and M. Maurer, "Ontology Based Scene Creation for the Development of Automated Vehicles", in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018

[14] F. Schuldt, "Ein Beitrag für den methodischen Test von automatisierten Fahrfunktionen mit Hilfe von virtuellen Umgebungen-English title: Towards testing of automated driving functions in virtual driving environments", Ph. D. dissertation, Technische Universität Braunschweig, 2017

[15] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles", in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1821–1827

[16] P. Junietz, F. Bonakdar, B. Klamann, and H. Winner, "Criticality Metric for the Safety Validation of Automated Driving using Model Predictive Trajectory Optimization", in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 60–65

[17] H. Abbas, M. E. O'Kelly, A. Rodionova, and R. Mangharam, "A Driver's License Test for Driverless Vehicles", *Mechanical Engineering Magazine Select Articles*, vol. 139, no. 12, S13-S16, 2017

[18] M. O'Kelly, H. Abbas, and R. Mangharam, "Computer-aided design for safe autonomous vehicles", in *Resilience Week (RWS), 2017*, 2017, pp. 90–96

[19] M. Althoff and S. Lutz, "Automatic Generation of Safety-Critical Test Scenarios for Collision Avoidance of Road Vehicles", in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1326–1333

[20] S. Hallerbach, Y. Xia, U. Eberle, and F. Koester, "Simulation-based identification of critical scenarios for cooperative and automated vehicles", *SAE International Journal of Connected and Automated Vehicles*, vol. 1, no. 2018-01-1066, pp. 93–106, 2018

[21] A. Pütz, A. Zlocki, J. Bock, and L. Eckstein, "System validation of highly automated vehicles with a database of relevant traffic scenarios", in *12th ITS European Congress*, 2017

[22] M. Grindal, J. Offutt, and S. F. Andler, "Combination testing strategies: a survey", *Software Testing, Verification and Reliability*, vol. 15, no. 3, pp. 167–199, 2005

[23] I. Sommerville, *Software engineering,* 8th ed.: New York: Addison-Wesley, 2006

[24] A. Arcuri and L. Briand, "Formal analysis of the probability of interaction fault detection using random testing", *IIEEE Trans. Software Eng.*, vol. 38, no. 5, pp. 1088–1099, 2012

[25] C. Amersbach and H. Winner, "Functional Decomposition - A Contribution to Overcome the Parameter Space Explosion during Validation of Highly Automated Driving (Accepted)", in *26th International Technical Conference and exhibition on the Enhanced Safety of Vehicles (ESV)*, 2019

[26] D. R. Kuhn, D. R. Wallace, and A. M. Gallo, "Software fault interactions and implications for software testing", *IEEE transactions on software engineering*, vol. 30, no. 6, pp. 418–421, 2004

[27] *ISO/PAS 21448:2019 Road vehicles - Safety of the intended functionality*, 2019

[28] F. Gao, J. Duan, Y. He, and Z. Wang, "A test scenario automatic generation strategy for intelligent driving systems", *Mathematical Problems in Engineering*, vol. 2019, 2019

[29] N. Kalra and S. M. Paddock, "Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?", 2016. [Online] Available: http://www.rand.org/pubs/research_reports/RR1478.html

[30] M. Holder *et al.,* "Measurements revealing Challenges in Radar Sensor Modeling for Virtual Validation of Autonomous Driving", in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2616–2622

[31] P. Cao, W. Wachenfeld, and H. Winner, "Perception sensor modeling for virtual validation of automated driving", *it-Information Technology*, vol. 57, no. 4, pp. 243–251, 2015

[32] M. Christiansen, *In geheimer Mission: auf Abnahmefahrt mit der neuen Mercedes E-Klasse, W213*. [Online] Available: http://5komma6.mercedes-benz-passion.com/in-geheimer-mission-auf-abnahmefahrt-mit-der-neuen-mercedes-e-klasse-w213/. Accessed on: Jul. 12 2017

[33] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems", in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2118–2125

[34] C. Amersbach and H. Winner, "Functional Decomposition: An Approach to Reduce the Approval Effort for Highly Automated Driving", in *8. Tagung Fahrerassistenz*, 2017