Validation of automated driving – a structured analysis and survey of approaches

Jan Erik Stellet, Matthias Woehrle, Tino Brade,
Alexander Poddey, Wolfgang Branz*

Zusammenfassung: Validating the safety of automated driving (AD) is a problem of remarkable complexity and practical relevance. New approaches are needed since a statistical proof of safety based on field testing does not scale. Despite the attention paid to this topic in industry and academia, a consensus or unified framework has not yet been reached. This work describes and compares four distinct validation approaches. Our findings reveal that the current fragmentary landscape can be partly explained by differences in the AD use cases. On the one hand, there are different problem spaces characterised, e.g., by the operational domain and the driving tasks. On the other hand, the solution space differs for business models related to end-customer vehicles and mobility as a service.

Schlüsselwörter: Automated Driving, Validation, Safety, SOTIF

1 Introduction

Although modern driver assistance systems (SAE L2 [1]) can temporarily take over control of the vehicle, they do not actually take over the responsibility which remains with the human driver. Thus, advancing to automated driving (AD) of SAE L3 and beyond poses much higher requirements for the development and validation of safe systems. Over the last years, this topic has received remarkable attention in industry and academia. However, there seems to be no consensus yet.

The principal challenges of validation of complex systems operated in an open context have been discussed and formalised in [2]: Firstly, acceptance criteria are situation-dependent and currently informal. Secondly, a real-world operational design domain (ODD) is unstructured and bears infinitely many possible interactions [3]. Thirdly, the emergent behaviours are hard to predict since they are the result of a complex interplay of components (possibly including machine learning).

In general, none of these three interdependent aspects can be formally expressed in a sufficiently complete manner. Therefore, there does not exist a standard notion of coverage in this domain. Additionally, a naive approach based on the enumeration of combinations of equivalence classes on relevant dimensions, e.g., road topology, dynamic objects and environmental conditions, results in exponential blowup and renders an exhaustive set of verification tests practically infeasible. Note that the primary concern lies beyond the

*The authors are with Robert Bosch GmbH, Corporate Research, Robert-Bosch-Campus 1, 71272 Renningen, Germany, firstname.lastname@de.bosch.com

scope of ISO 262626 [4] which does not define the necessary *'nominal performance'* of a safe system. The complementary standard ISO PAS 21448 [5] addresses *'functional insufficiencies'* but currently does not provide a detailed strategy on how to identify them.

In this research work, we compare prominent proposals for AD safety validation and highlight their similarities and differences. The goal is to help the reader understand the different approaches and inherent challenges. Thereby, we want to initiate a discussion in the academic community about research directions related to validation of AD.

This paper is organised as follows: First, related work will be discussed (Sec. 2). Thereafter, a formal problem description will be introduced in Sec. 3. Based on this, Sec. 4 will analyse four proposals. The paper concludes with Sec. 5.

2 Related work

Our previous publication [6] surveys the state of the art of testing of driver assistance systems (SAE L1–2) at that time. The present work provides an extended and updated view with a focus on automated driving systems (SAE L3–5).

Concerning the development of AD systems, [7] identifies and describes challenges from both sides of the V-model process, e.g. challenges related to fail-operational designs and complex requirements as well as testing of non-deterministic algorithms. The work [8] discusses stakeholder (e.g., legislative) perspectives on testing of AD systems. Three method clusters are identified, namely real world driving, formal verification and scenario-based tests. The test methods are categorised based on the representation of the object under test, the stimulus and assessment criteria.

Complementing the existing work, we identify practical AD validation challenges that allow us to differentiate among validation approaches.

3 Problem description and challenges

The question how to validate an AD system via testing procedures has become more and more important. Due to the more complex problem, more advanced methods than a representative field test (black-box test) as employed for validation of driver assistance systems are needed. Thereby, several practical challenges have to be addressed:

C1: **Representativeness challenge:** The goal of a test is to predict properties of the realisation in its operational domain. Therefore, a set of situations has to be sampled in order to accurately predict the operational context with respect to the purpose [11].

However, the openness of the context means that enumeration of situations does not scale. Instead, a global random sampling strategy can be used to obtain a representative set of situations, as it is done in field test drives. But, since such an undirected sampling will encounter (small) critical subspaces only with a certain (small) probability, the test duration increases with a decreasing frequency of critical events.

These considerations are illustrated by the empirical statistics on disengagements of autonomous vehicles [9]. Two examples, visualised in Fig. 1, show how the rate of

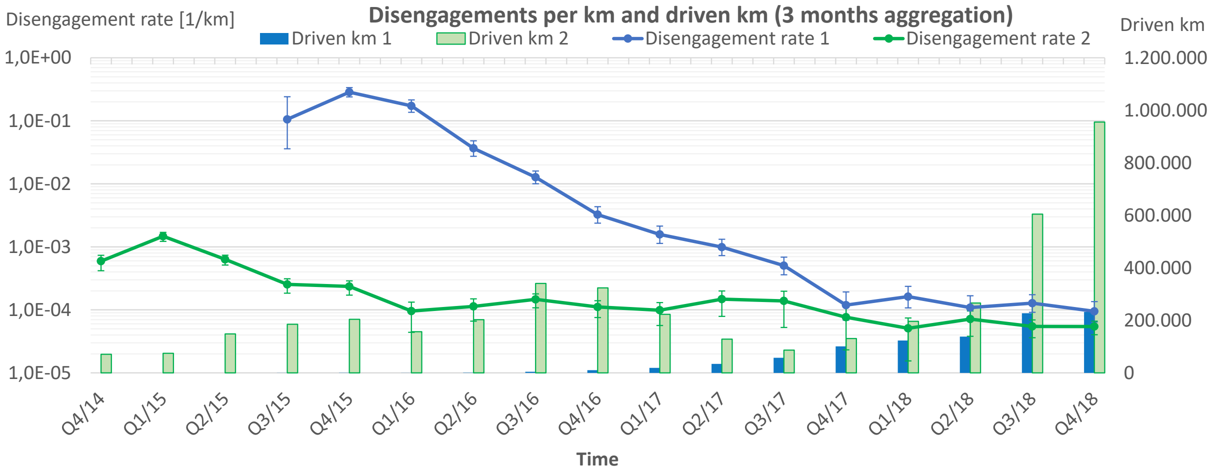

Figure 1: Two examples of human driver intervention (*disengagement*) rates and driven kilometres as published at [9]. Note that it is generally not known if all data is comparable because the driving profile is unknown and might change over time, e.g. towards more challenging situations. Further discussions can be found in [10].

improvement reduces over time although testing distances are significantly increased. This could be regarded as a first indication for a *'heavy tail safety ceiling'* [12].

C2: **Closed-loop challenge:** The AD function influences the vehicle's trajectory and thus the sensors' perspective of the world. Moreover, due to interactions, the future trajectories of other traffic participants are also influenced by the automated vehicle's behaviour. Thus, assessing an AD system requires to observe it in closed loop, at least partially to consider the interaction with the environment.[1]

C3: **Modelling challenge:** Testing in a virtual environment requires that the context (including e.g. behaviour of other traffic) and the realisation (including e.g. environment sensors such as cameras) have to be modelled. Additionally, the validity of the models has to be shown, see [11] for a detailed formalisation of this challenge.

C4: **Continuous improvements challenge:** Due to the complexity of the task the development of an AD function is necessarily iterative. However, each modification may change the behaviour of the automated vehicle and the interaction with others (cf. closed-loop challenge) compared to a recorded situation. Thus, conclusions from previous testing data as well as models may become invalid.

C5: **Insights challenge:** Testing on system level, i.e. in a black-box manner without exploiting structure in the realisation, prevents insights on sub-critical failures (cf. [13, 14]). Thus, this kind of testing becomes largely uninformative and inefficient.

C6: **Cross-domain, system-of-system challenge:** End-to-end performance relies on several, interacting systems from different domains in sensing, planning and actuation that rely on different design and validation approaches.

[1] Note that this is different for certain driver assistance systems, such as an automatic emergency brake, where solely the occurrence of a false intervention of an otherwise passive system shall be validated.

C7: **Supervision challenge:** Supervising an active automated driving function during a real-world driving test is a very demanding task for humans, especially if necessary interventions are very rare.

Note that there are additional challenges that are common to all approaches, such as legal aspects and the safety of data-based algorithms (e.g. machine learning) [7].

4 Survey of validation approaches

4.1 Passive AD (shadow mode)

The basic idea of the *shadow mode* or *passive AD* approach is to use a large fleet of human-driven vehicles to collect raw sensor data. After these open-loop recordings are retrieved, a simulation environment is used for a closed-loop replay of the data to the AD function [15, 16]. In another variant, the simulation environment runs on-line in the vehicle and only some data, determined by trigger conditions, is retrieved [17].

Since the AD function is passive during the data collection phase, it exhibits no safety risk. This enables a wide-spread roll-out to many end-user vehicles and would allow achieving large test distances in manageable time. Technical challenges for implementing the approach are:

- A sufficiently complete and correct transfer of the vehicle's environment to the simulation environment is necessary. Additional reference sensors, sensor data post-processing or annotations by humans can be used [17], but the correctness and completeness needs to be validated.
- For the simulation environment, the behaviour of other traffic participants has to be modelled, in order to close the loop.
- If trigger conditions are used to identify and transmit only relevant data from critical situations, these triggers have to be validated for sufficient completeness and correctness. Otherwise, a situation where the AD function would have planned a dangerous trajectory could remain unnoticed.
- The transferability of data collected with human-controlled vehicles to autonomously driven ones has to be argued. For example, it is sensible to assume that an AD vehicle drives defensively, e.g. leaving sufficient gaps, which might provoke more cut-in situations than a human driver would experience.

Conclusion Silently testing in end-customer vehicles improves scalability since there is no supervision challenge (C7) and the passive function can be continuously improved in the background (C4). However, there are several caveats that must be addressed before the open-loop recordings can be used as a validation argument. On the one hand, a sufficiently accurate simulation environment is needed to close the loop (C2). The dependence on models (e.g. reference environment model, sensor and behaviour models) reinforces the modelling challenge (C3). On the other hand, the representativeness of the data from human-controlled vehicles with respect to the AD function needs to be argued (C1). Additionally, if data can be recorded only selectively, there is limited insight into the system (C5).

4.2 Formally safe planning and statistically validated sensing

The authors of [18][2] argue that scalability in the sense of mass production and *'everywhere'* automated driving is not feasible with merely statistical data-driven validation. As a solution, they propose to combine a data-driven validation of the perception system and a formal model that guarantees for the safety of planned trajectories given a correct environment perception.

Data-driven validation of the perception system Firstly, the rate of situations being erroneously considered unsafe (*safety-critical ghosts*) or erroneously considered safe (*safety-critical misses*) shall be statistically validated by field tests.[3] Multiple parallel sub-systems, preferably based on different sensor technologies, with an assumed known probability of common errors are used in order to argue a lower failure rate of the combined perception system. Technical challenges with this approach are:

- Assumptions about the probability of common errors have to be validated. Note that small errors can cause a strong underestimation of the overall failure rate [19].
- Reference (ground truth) data of the environment model is needed in order to estimate the rates of safety-critical ghosts and misses. A straight-forward calculation based on the assumptions in [18] necessitates reference data on the order of magnitude of 10^5 h.
- Estimating the frequency of ghosts and misses is a not symmetric problem. While ghosts may occur at any time, a safety-critical miss can only occur if there is a dangerous situation. This imbalance may lead to much wider confidence intervals (or increased test durations) when estimating the frequency of safety-critical misses.
- Similar to the approach from Sec. 4.1, the profile of data collected for perception validation might differ from the statistics if the vehicles drive autonomously.

Formal guarantees of the safety of the planned trajectories The planning component is designed to be intrinsically safe by means of a safety envelope. To this end, the Responsibility Sensitive Safety (RSS) model is introduced. This model aims for a universally valid and explainable rule set that confines the actions of the actual trajectory planner. To achieve this, the *'elusive directive called duty of care'* is made explicit in form of five rules, e.g. *'right-of-way is given, not taken'* [18]. From these rules, quantitative constraints, e.g. in the sense of minimum safety distances, are derived.

Achieving generally accepted rules relies on two important premises. Firstly, the set of traffic scenarios used to derive the rules are exhaustive, especially when it comes to exceptional situations [3]. Secondly, the behaviour of other traffic participants can be

[2] Note that we directly refers to v6 of [18], as there have been substantial changes in the past.

[3] Note that there is an important distinction between *safety-critical ghosts* or *misses* and *false positive* or *false negative detections* in the perception system. Although detection errors are the usual causes for the former, a ghost or miss can also be caused by measurement errors, e.g. a noisy distance measurement. Furthermore, not every false positive or negative detection of the sensors will change the judgement of a situation as safe or dangerous, i.e. produce a safety-critical ghost or miss, respectively. Therefore, these error definitions are tied to the formulation of safety envelope which decides whether a situation is safe or dangerous.

modelled with a realistic set of parameters, e.g. the maximum reasonable deceleration that a lead vehicle might apply. It is currently unclear how realistic parameter values can be obtained or if it is even possible to assign single values that are adequate in all situations. Instead, it might be necessary to let the parameter values depend on the current situation to represent the societally accepted boundary between agile and dangerous behaviour. Implementing a rule set requires modelling of behaviour and motion of the ego vehicle and other traffic participants. Shalev-Shwartz *et al.* [18] use simple kinematic constraints, however more general techniques such as reachability analysis provide stronger guarantees and allow integrating further sources of uncertainty [20].

Conclusion Overall, the idea of formalising acceptable behaviour of an AD vehicle has its strengths in formally resolving the closed-loop (C2) and supervision challenges (C7). Moreover, explainable rules can help in increasing insight and transparency of the behaviour planning (C5). The cross-domain challenge (C6) is addressed by separation of concerns between perception and planning. However, to be applicable, error probabilities of the perception system have to be estimated and a generally valid set of parameters in the RSS model has to be argued. Since the model and its parametrisation take a central role, the corresponding modelling challenge (C3) is a key challenge for this approach.

4.3 PEGASUS: Scenario-based (top-down) approach

PEGASUS[4], a German publicly funded project, aimed at developing methods for ensuring the safety of automated driving on the example of a SAE L3 (conditional automation) *'highway chauffeur'* function.

The core of PEGASUS approach are scenarios. Their description is based on a six-layer model to compositionally model static and dynamics aspects in a joint description [21]. Scenarios can be identified from system knowledge [22], domain modelling [21,23] and field observation [17]. Scenarios feature parameters that can be varied in order to increase their coverage.

Scenarios are used as test cases that are executed and evaluated in simulation or on test tracks. Corresponding criticality metrics are used to determine the automated driving capabilities. Virtual testing enables reproducible tests and large scale parameter variations. One goal of test track testing is a point-wise validation of the simulation model. Additionally, real-world drive tests are conducted.

The scenario-based approach is developed together with an assessment of its validity and two major limitations have been identified:

1. Risk of generating the wrong scenarios: On the one hand, if scenarios are derived deductively the completeness and relevance is difficult to achieve. On the other hand, if scenarios are identified inductively from field data, the completeness depends on the metrics and models used. Additionally, the data might be incompletely mapped to a test case.

2. Risk of a wrong selection (reduction) of scenarios: Test cases might be defined based on equivalence classes although the underlying scenarios are in fact not equivalent.

[4] Project for the Establishment of Generally Accepted quality criteria, tools and methods as well as Scenarios and Situations

Concerning scenario parameters, the challenge is to find representative value ranges, a suitable discretisation and to cope with the exponentially increasing number of combinations.

Conclusion The PEGASUS approach attempts to standardise and make AD safety requirements transparent with a scenario-based approach. The most challenging aspect of this is to find the right set of scenarios (C1 representativeness). The closed-loop (C2) and insight (C5) challenges are alleviated by also using simulations for test case execution. However, before this can be achieved, the simulation environment has to be validated (C5 modelling). The PEGASUS concept addresses this with a cross-check between test cases executed on test tracks and in simulation. Nevertheless, representativeness and modelling are core challenges for the PEGASUS approach.

4.4 Continuous validation

The previously described validation approaches mostly break down the overall problem into several successive steps, e.g. sensor data is collected and the safety of the AD system argued by simulation (Sec. 4.1) or statistical analysis and a formal model (Sec. 4.2).Thus, there are expectations that some elements, e.g. a simulation environment, will have been developed and validated because other elements rely on them.

However, this may be problematic in practice due to dependencies and a necessarily iterative development of an AD system. An alternative is to consider the development of the AD system and the validation infrastructure as a joint iterative process. An applied example with the scope of validating SAE L4 robotaxis can be found in [24].

Iterative data- and simulation-driven development The core of this approach is an iterative development and testing cycle with strong links between the different test strategies as illustrated in Fig. 2. This includes domain and system analyses, test track testing, virtual testing and field tests. The complementary nature is key to efficiency and effectiveness of the iterative cycle.

Exemplarily, field testing can identify edge case scenarios that are hard to identify otherwise. However, to cope with the continuous improvement challenge, recorded field test data must be reusable in a closed-loop simulation environment for reproducibly testing and generalising such scenarios. In addition to the random sampling-based field tests, test cases should be derived from systematic analysis of both the AD system and the simulation environment. Improving the realism of the simulation environment is equally important in order to rely on the efficiency of virtual testing. Otherwise, critical scenarios might be overlooked in virtual testing and would turn up later as surprises in field tests.

Incremental deployment and supervised operation A mobility as a service application benefits an iterative approach by means of the operation and deployment strategy. Firstly, the service can be launched under supervision of a safety driver. Secondly, a fine-grained and incrementally growing ODD is possible. The ODD for the fully automated vehicles (i.e. without safety driver) within a larger fleet can be defined on the level of individual customer trips. Third, frequent maintenance stops and remote monitoring of all vehicles are possible.

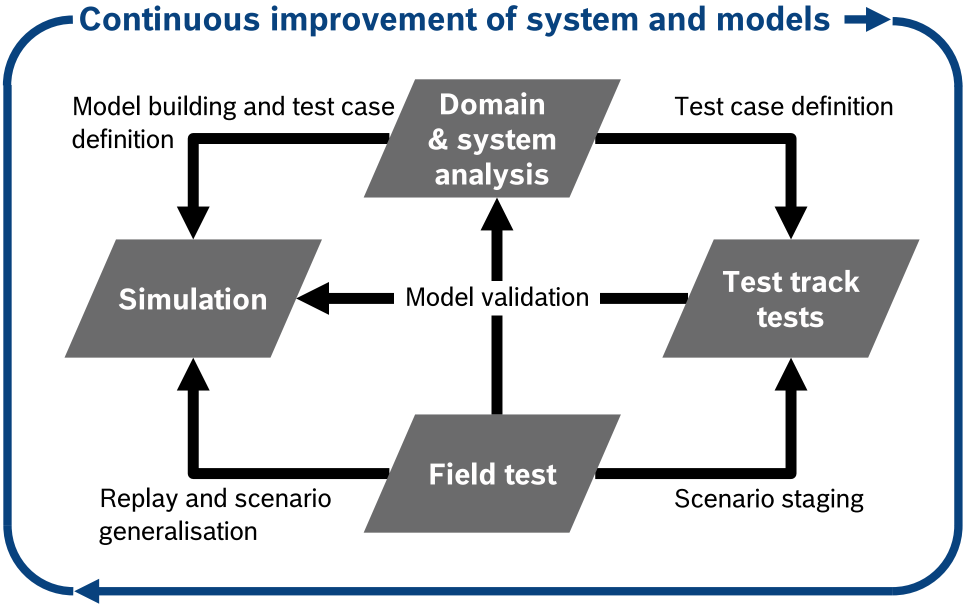
Continuous improvement of system and models
Model building and test case definition
Domain & system analysis
Test case definition
Simulation
Model validation
Test track tests
Field test
Replay and scenario generalisation
Scenario staging

Figure 2: The iterative approach builds on continuous improvements of the system and models in a feedback loop with multiple ways to generate test cases.

Conclusion The sketched approach relies on similar elements as the ones discussed before, but integrates these complementary elements tightly. An iterative approach benefits from an incremental deployment that is only possible in mobility as a service applications. However, it is hampered with respect to scalability as new service locations are iteratively added and may exacerbate the continuous improvement challenge (C4), especially in the long tail, as new location may invalidate previously collected data, knowledge and models. This depends heavily on the balance between implicit knowledge such as field-data and explicit knowledge such as perception models including common errors. The supervision challenge (C7) must be specifically considered as it is difficult for humans to supervise an ever more capable automated driving function. Furthermore, the iterative nature must not conceal the fact that eventually, acceptance depends on explainable criteria. Thus, the approach could benefit from elements found in other, e.g. transparent rules for safe behaviour or coverage criteria.

5 Conclusion

As we have seen, different AD use cases, e.g. highway chauffeur or robotaxi, can induce quite different safety validation concepts. One reason is that the associated business case has a large impact on validation and its challenges, e.g., by carefully restricting the ODD or alleviating the supervision challenge. This concerns the initial release scope and its corresponding ODD but also scalability with respect to vehicle variants, geographic distribution and resulting diversity in ODDs. On the one hand, OEMs and suppliers of end-customer vehicles have to consider many vehicle variants in different price ranges that are used worldwide. On the other hand, driver-less mobility services are typically characterised by a homogeneous vehicle fleet and a restricted operational domain. The vehicle fleet might initially consist of vehicles with and without a supervisor, enabling a fine-grained and growing ODD.

This work provides some first insights by identifying and discussing several validation challenges. While overall metrics and information for a public audience are available for individual approaches, detailed information to further analyse the progress with respect to the presented validation challenges is currently lacking. In future work, we try to reduce this information gap by further detailing on validation challenges and how current and future approaches could address them.

References

[1] SAE International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," 2018.

[2] J. E. Stellet, T. Brade, A. Poddey, and W. Branz, "Formalisation and algorithmic approach to the automated driving validation problem," in *Intelligent Vehicles Symposium (IV), IEEE*, pp. 45–51, 2019.

[3] P. Koopman and F. Fratrik, "How many operational design domains, objects, and events?," 2019.

[4] International Organization for Standardization, "Road vehicles – functional safety," 2011.

[5] International Organization for Standardization, "Road vehicles – safety of the intended functionality," 2019.

[6] J. E. Stellet, M. R. Zofka, J. Schumacher, T. Schamm, and J. M. Zöllner, "Testing of advanced driver assistance towards automated driving: A survey and taxonomy on existing approaches and open questions," in *Intelligent Transportation Systems (ITSC), 18th IEEE International Conference on*, pp. 1455–1462, 2015.

[7] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE International Journal of Transportation Safety*, vol. 4, no. 1, pp. 15–24, 2016.

[8] P. Junietz, W. Wachenfeld, K. Klonecki, and H. Winner, "Evaluation of different approaches to address safety validation of automated driving," in *21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 491–496, 2018.

[9] "Autonomous vehicle disengagement reports," tech. rep., State of California: Department of Motor vehicles.

[10] C. Lv, D. Cao, Y. Zhao, D. J. Auger, M. Sullman, H. Wang, L. M. Dutka, L. Skrypchuk, and A. Mouzakitis, "Analysis of autopilot disengagements occurring during autonomous vehicle testing," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 1, pp. 58–68, 2018.

[11] E. Böde, M. Büker, U. Eberle, M. Fränzle, S. Gerwinn, and B. Kramer, "Efficient splitting of test and simulation cases for the verification of highly automated driving

functions," in *Computer Safety, Reliability, and Security* (B. Gallina, A. Skavhaug, and F. Bitsch, eds.), (Cham), pp. 139–153, Springer International Publishing, 2018.

[12] P. Koopman, "The heavy tail safety ceiling," in *Automated and Connected Vehicle Systems Testing Symposium*, 2018.

[13] S. Shalev-Shwartz and A. Shashua, "On the sample complexity of end-to-end training vs. semantic abstraction training," *arXiv preprint arXiv:1604.06915*, 2016.

[14] H. Winner, W. Wachenfeld, and P. Junietz, *Validation and Introduction of Automated Driving*, pp. 177–196. Cham: Springer International Publishing, 2018.

[15] W. Wachenfeld and H. Winner, "Virtual assessment of automation in field operation a new runtime validation method," in *Workshop Fahrerassistenzsysteme, Walting, Germany*, pp. 161–170, 2015.

[16] A. Koenig, K. Witzlsperger, F. Leutwiler, and S. Hohmann, "Overview of had validation and passive had as a concept for validating highly automated cars," *at-Automatisierungstechnik*, vol. 66, no. 2, pp. 132–145, 2018.

[17] P. Junietz, W. Wachenfeld, V. Schönemann, K. Domhardt, W. Tribelhorn, and H. Winner, "Gaining knowledge on Automated Driving's safety – the risk-free VAAFO tool," in *Control Strategies for Advanced Driver Assistance Systems and Autonomous Driving Functions*, pp. 47–65, Springer, 2019.

[18] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374v6*, 2017.

[19] R. W. Butler and G. B. Finelli, "The infeasibility of quantifying the reliability of life-critical real-time software," *IEEE Transactions on Software Engineering*, vol. 19, no. 1, pp. 3–12, 1993.

[20] M. Althoff and J. M. Dolan, "Online verification of automated road vehicles using reachability analysis," *IEEE Transactions on Robotics*, vol. 30, no. 4, pp. 903–918, 2014.

[21] F. Schuldt, *Ein Beitrag für den methodischen Test von automatisierten Fahrfunktionen mit Hilfe von virtuellen Umgebungen.* PhD thesis, Technische Universität Braunschweig, 2017.

[22] M. Büker, B. Kramer, E. Böde, S. Vander Maelen, and M. Fränzle, "Identifikation von automationsrisiken hochautomatisierter fahrfunktionen in pegasus," in *AAET 2019 – Automatisiertes und vernetztes Fahren*, 2019.

[23] G. Bagschik, T. Menzel, and M. Maurer, "Ontology based scene creation for the development of automated vehicles," in *IEEE Intelligent Vehicles Symposium (IV)*, pp. 1813–1820, 2018.

[24] "Waymo safety report: On the road to fully self-driving," tech. rep., Waymo LLC, 2018.