electronics

MDPI

Article

Explainable Safety Argumentation for the Deployment of Automated Vehicles

Patrick Weissensteiner * and Georg Stettinger

Infineon Technologies AG, 85579 Neubiberg, Germany; georg.stettinger@infineon.com
* Correspondence: patrick.weissensteiner@infineon.com

Abstract: With over 1.6 million traffic deaths in 2016, automated vehicles equipped with automated driving systems (ADSs) have the potential to increase traffic safety by assuming human driving tasks within the operational design domain (ODD). However, safety validation is challenging due to the open-context problem. Current strategies, such as pure driving and requirement-based testing, are insufficient. Scenario-based testing offers a solution but necessitates appropriate scenario selection, testing methods, and evaluation criteria. This paper builds upon a method to calculate the covered ODD using tested scenarios generated from logical scenarios, considering parameter discretisation uncertainty. Acceptance criteria for the safety argumentation are proposed based on parameter space coverage and variance introduced via discretisation, thus contributing to quantifying the residual risks of safety validation. The approach is demonstrated through two logical scenarios with probability density functions of the parameters generated using a trajectory dataset. These criteria can serve as risk acceptance criteria, providing comparability and explainable results. By developing a robust scenario-based testing approach, ADS safety can be validated, leading to increased traffic safety and reduced fatalities. Since ADSs incorporate AI models, this proposed validation strategy can be extended to AI systems across multiple domains for the respective assurance argument required for deployment.

Keywords: automated vehicles; autonomous vehicles; automated driving; validation; coverage; operational design domain; automotive safety; vehicle safety; safety argumentation

Citation: Weissensteiner, P.; Stettinger, G. Explainable Safety Argumentation for the Deployment of Automated Vehicles. *Electronics* **2024**, *13*, 4606. https://doi.org/10.3390/electronics13234606

Academic Editors: Alexander Gegov, Raheleh Jafari, Femi Isiaq and Kalin Penev

Received: 30 October 2024
Revised: 16 November 2024
Accepted: 20 November 2024
Published: 22 November 2024

1. Introduction

In 2016, more than 1.3 million traffic deaths happened worldwide [1]. Globally, the death rate has stagnated whilst the overall population is still growing. To some degree, this is attributed to the increased safety measures in vehicles. In particular, the European Union (EU) has reduced deaths due to traffic accidents in recent decades [2]. This reduction is primarily due to the advent of advanced driver assistance systems (ADASs) in modern vehicles, which take over part of the human driving task. In contrast, automated vehicles (AVs) are cyber–physical systems equipped with an automated driving system (ADS), taking over the complete driving task of the human in the target operational domain (TOD) [3]. The TOD represents the area of ADS deployment and can include certain conditions outside of the operational design domain (ODD) of the ADS. The ODD is used to describe the boundaries of the ADS-equipped vehicle and the conditions under which the ADS is designed to operate, including environmental, geographical, and time-of-day aspects, as well as potential restrictions [4]. The differences between the TOD and the ODD of an ADS highlight the limitations of such systems [5]. In addition, the respective AV behaviour is defined by the behaviour capabilities [6].

In any case, ADS-equipped vehicles operate in an open-context environment, which cannot be described during the system's design time fully and which constantly changes over time [7,8]. Despite these difficulties, the potential benefits of AVs are manifold. The expected benefits range from ecological and economic benefits to increased accessibility,

comfort and traffic safety. Overall, global megatrends such as demographic and societal changes, rapid urbanisation, and climate change are influencing the evolution of smart mobility, with shared AVs being one of the core elements [9,10].

The aforementioned open context, in addition to other challenges in the respective ODD (such as complex urban environments, including the behaviour of other traffic participants, such as pedestrians), presents enormous difficulties, especially for the perception system of such vehicles. Despite recent advancements in real-time perception [11], the safety validation of the complete ADS remains challenging [12]. Different validation strategies to overcome these challenges are discussed in [13,14]. Section 1 presents a detailed discussion of current challenges in the safety validation for ADS-equipped vehicles.

Challenges of Safety Validation for ADS

Due to the mentioned open-context environment, ADS-equipped vehicles are deployed, and various challenges appear. On the one hand, technical failures can occur due to (random) software- or hardware faults. ISO26262 [15] tackles such functional safety aspects. On the other hand, ensuring the system's intended functionality is safe is tackled in the safety of the intended functionality (SOTIF) standard [16]. The SOTIF standard requires the safety assessment of ADS-equipped vehicles. Previously, a function-based approach was the standard procedure for ADAS functions. This approach means that pre-defined tests to confirm the functionality (e.g., UN ECE R131 for advanced emergency braking systems) are defined. However, this approach is unsuitable for AVs, as it would lead to performance optimisation based on the chosen tests and would not accurately determine an the safety performance of an AV in its ODD. Hence, another approach emerged for more complex ADASs and for ADSs in general, which is scenario-based testing (SBT). It is based on the scenario concept introduced in [17] and extended in [18]. Various standardisation activities are ongoing considering the usage of the scenario concept for the development, testing, and validation activities of AVs [19].

Based on the concept of scenarios, there exist various validation strategies [20–22]. The two main techniques to distinguish are falsification- and testing-based approaches. Falsification-based approaches are trying to find counterexamples violating the safety requirements. However, such approaches cannot provide general statements about the AV performance in the ODD but do serve a purpose for the efficient development of ADSs [20]. The testing-based approach, as defined in [20], determines if the requirements regarding safety are satisfied based on a finite set of test scenarios. In principle, testing-based SBT can provide statements about the AV performance in the respective ODD if an unfeasible number of scenarios is executed [20]. Hence, efficient methods for selecting scenarios and quantifying their contribution towards the coverage of the respective scenario space are required. In addition, approaches for interpreting the scenario space coverage regarding the target ODD in an explainable manner are necessary [23].

Overall, the challenges of SBT regarding safety validation concern the selection of scenarios, test methods, and the actual assessment of the testing results at the microscopic and macroscopic levels (see Figure 1):

- Scenarios: If used for AV safety validation using SBT, scenarios must be generated to uncover as many unknown hazardous situations as possible, in line with the SOTIF principle. Based on [24], such scenario generation techniques can be data driven (e.g., using datasets [25]), optimisation based, combinational, or expert based (e.g., [18,26]). Combinations of these approaches are also possible. Usually, various continuous parameters (CPs) define a logical scenario (LS), which, after discretisation, leads to concrete scenarios (CSs) that can be tested. The coverage-based testing method (see Section 4), used as a basis for the risk acceptance criteria in Section 5, utilises a combination of scenario generation techniques. Concretely, the method combines a statistical data-driven and combinational approach. Based on the analysis of [24], these two approaches complement each other concerning hazardous situations.

- Test methods: The scenarios are tested using different test methods, ranging from virtual to real-world approaches. An overview is given in [27].
- Safety assessment: The actual safety assessment can be split into microscopic and macroscopic assessments. In the microscopic assessment, the individual scenarios used to test the ADS are evaluated using respective metrics [20,28]. In the macroscopic assessment, (mostly statistical) statements about the overall impact on AVs can be made [20,29]. The macroscopic statement must serve as an essential argument for introducing AVs into real-world traffic. For example, such a statement can be achieved by providing proof of a lower accident probability than human drivers.

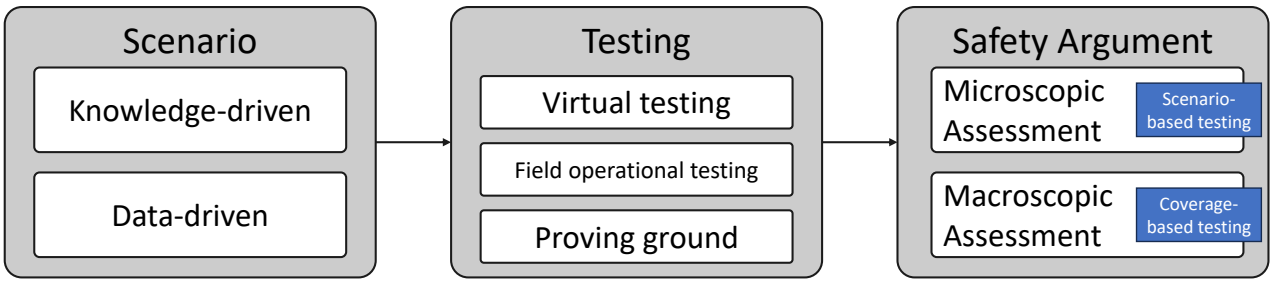

Figure 1. The different stages of SBT are displayed, from the scenarios and the testing methods to the approaches towards safety argumentation. Adapted from [20,30].

SBT is already utilised in various large research projects developing scenario-based methodologies in Germany [31], Japan [32], and across the EU [33]. Furthermore, the United Nations Economic Commission for Europe (UNECE) proposed a regulatory framework for the safety assessment of ADSs [34,35]. Moreover, the UNECE WP.29 Working Party on Automated/Autonomous and Connected Vehicles developed the New Assessment/Test Method (NATM) to certify AV. It contains a multi-pillar approach towards demonstrating a valid safety assessment of an ADS by incorporating pre-market deployment and in-service operation [36].

Although SBT is considered the most promising option for AV safety assurance, additional aspects are needed for mixed-traffic environments. As the penetration rate of AVs increases, impacts regarding traffic flow and other metrics are expected [37]. In addition, greater AV penetration could lead to greater exposure to critical events, which could dilute the positive effects of AV deployment [38]. This could lead to changes in the distribution of scenario parameters, which needs to be considered going forward. Furthermore, an increasing AV penetration rate needs to be considered in the scenario design, using an SBT strategy and including behaviour models in simulation—not related to human driving behaviour but to an AV.

Overall, SBT is widely acknowledged as one of the critical aspects of ADS safety validation. However, there is currently a gap between assessing the capabilities of ADS-equipped vehicles on a microscopic level and deriving an actual macroscopic statement concerning the targeted ODD in an explainable manner. Hence, this paper proposes an advancement in that regard, building on top of the presented concept of ODD coverage in [23].

2. Structure of the Article

The remainder of the article is structured as follows. First, Section 3 introduces the relevant aspects to consider for the safe deployment of ADS-equipped vehicles, namely, the requirement of a compelling safety case. Second, in Section 4, various ways to define coverage—a term used in many domains for different aspects—are discussed. The article then uses the definition provided in [23]. Furthermore, the process of coverage determination based on [39] is discussed and used as the basis for the subsequent sections, as it

details ways to identify, select, and test scenarios based on a set of requirements, including the respective target ODD of an AV.

Third, Section 5 presents novel risk acceptance criteria as part of the safety argumentation necessary for any ADS-equipped vehicle deployment. Next, Section 6 presents the results of an application example. Section 7 discusses the generated results and their implications. The article concludes with a summary and an outlook in Section 8.

3. Safe Deployment of ADS-Equipped Vehicles

For the safe deployment of ADS-equipped vehicles, there needs to be an absence of unreasonable risk (AUR). This is an often-used definition of safety also used in ISO 26262 [15]. To achieve and argue for an AUR, a safety case—a structured argument supported by evidence—needs to be provided [40]. A goal-structuring notation (GSN) is often used to construct such a safety case. Essentially, respective safety evidence is utilised to form a safety argument to meet the overall safety requirements and objectives, the top goals formulated in the GSN [41]. Aurora provides a publicly available example of a safety case for an AV [42]. Consequently, to achieve and argue for AUR, the global top goal has to be safe behaviour in the target ODD [43], which includes functional safety [15] and cybersecurity as sub-goals, amongst others. Another sub-goal is the AUR, specifically due to insufficient performance and robustness regarding the intended functionality of the ADS-equipped vehicle in the target ODD—required by the SOTIF standard [16]. To showcase that the risk across the entire ODD, including the behaviour capabilities of the AV, is below an acceptable level (and therefore arguing for AUR) requires quantifying the residual risk levels in the target ODD.

In the safety literature, many concepts exist that can be used to argue for the AUR:

- ALARP (as low as reasonably practicable) [44] describes reducing the tolerable risk until negligible. However, many other interpretations exist (see [45]).
- GAMAB (globalement au moins aussi bon (translation: as a whole at least as good as) ([44], Annex D) would define, in the ADS context, that the introduced AV is at least as safe as the state of the art in current road traffic.
- MEM (minimum endogenous mortality) sets a threshold based on the rate of fatalities per operational metric (e.g., hours of operation) [45].
- Another type of concept is the notion of positive risk balance (PRB) [46]. It is a quantitative safety measure introduced by the German Ethics Commission and was further reworked in the informative ISO Technical Report 4804 [47,48]. However, the concept of PRB can be interpreted in several ways as stated by [45], which provides two potential interpretations. The first interpretation places PRB as the high-level goal in the overall assurance process. The second interpretation describes PRB as a method to determine tolerable risk, which in turn allows the use of it as a criterion for unacceptable risk and is subsequently used in [49].

Blumenthal et al. interpret safety as a process, threshold, and measurement [50]. The consideration of safety as a process is necessary for the overall system design of an ADS. However, for the SOTIF-related sub-goal of the overall safety case, using safety as a threshold and measurement to determine if the ADS-equipped vehicle behaviour can be carried out with the AUR is equally essential. Concretely, this must be determined before and after the deployment (pre- and post-deployment phase). This notion of pre- and post-deployment phases is similar to the concept of the AI life cycle [51,52], which contains the phases design, develop, and deploy (for the pre-deployment), and operational use, monitor, and evaluate and analyse (for the post-deployment).

It is crucial to consider if all of the above metrics are leading or lagging metrics [16,50] (also known as primary or secondary metrics in the AI domain). Leading metrics are evaluated using measures available in the pre-deployment phase, such as evaluation results from various test methods, from simulation to test tracks. Lagging metrics are evaluated using measures derived from statistical data gathered in the post-deployment phase. Figure 2 showcases this based on the ADS life cycle. Many of the mentioned metrics

to determine the AUR are much more accurate when implemented as lagging metrics. To implement metrics such as the MEM as a leading metric would require estimation, as the actual probability for failure is unknown for a specific ADS in a respective target ODD, which introduces error potential in the overall safety argumentation [22]. Hence, to make an informed deployment decision, safety validation strategies that explore the whole target ODD and respective scenario space and can determine certain residual risks before the actual deployment are required.

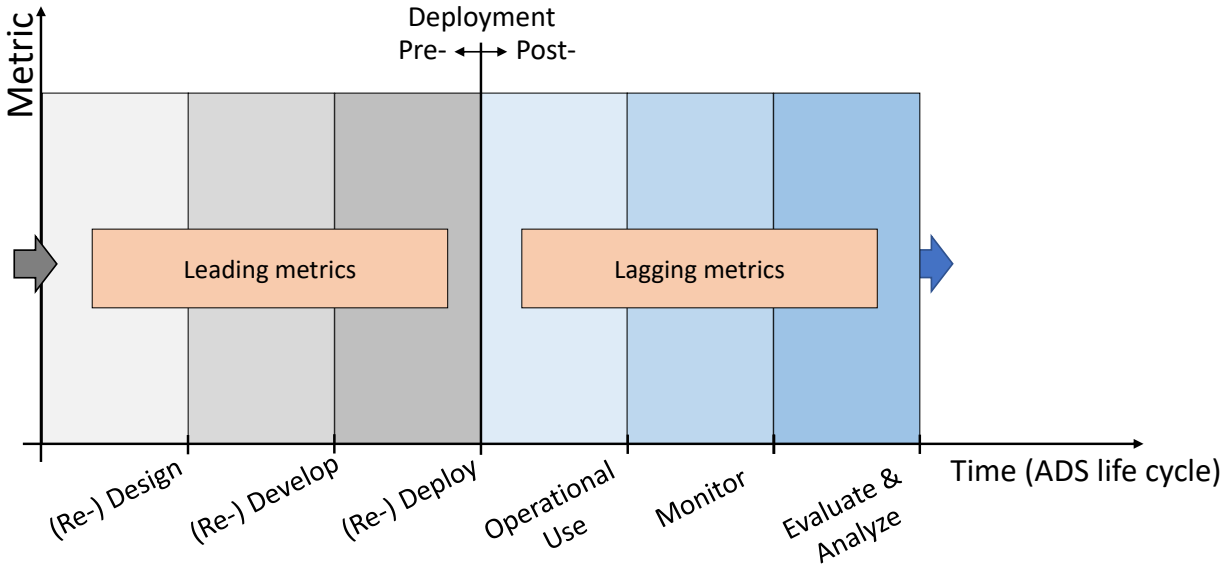

Figure 2. Leading metrics can be utilised in the pre-deployment phase, whereas lagging metrics are effective in the post-deployment phase.

4. Coverage-Based Testing Method

As explained in the previous section, a PRB is more straightforward to prove when using lagging metrics. However, such traditional distance-based statistical approaches, which can provide lagging metrics [22], are not available before or at the decision point for a specific AV deployment without using vague assumptions that can significantly deviate from the actual safety within the target ODD. Therefore, PRB and, most importantly, the AUR need to be proven additionally by assessing the safety performance of the ADS-equipped vehicle within the target ODD in the pre-deployment phase.

Hence, methods for scenario-space exploration are required. In [53], potential methods for identifying critical scenarios are provided, essentially using a guided-search or naive-search approach, mostly sampling or combinatorial testing. Another way of interpreting these two options is to see the guided search as a way to determine adversarial scenarios and the naive search as coverage-based testing. The guided-search approaches are based on optimisation regarding the search and identification of critical or adversarial scenarios rather than covering the scenario parameter space [54]. Another method for criticality identification to provide coverage estimates is proposed in [55]. However, certain assumptions about the behaviour of the ADS-equipped vehicle are made, which are difficult to sustain if a complete, production-ready ADS should be validated in terms of safety. Overall, the objective of the safety validation process is not to seek critical scenarios or specific edge cases (e.g., [56,57]). This search is crucial during these systems' development and internal validation, particularly across various design iterations, as it can impact certain development decisions. However, these edge cases inherently rely on the performance of the tested ADS concerning the experienced criticality based on the pre-defined metrics.

Overall, it needs to be mentioned that different test and scenario generation strategies can tackle different aspects of AV safety assessment in the context of SOTIF [24]. However, coverage-based testing appears to be the most promising option for generating evidence of AV safety across the defined target ODD. Concretely, the strategy of coverage-based testing is to execute concrete tests to cover the scenario space [54]. Furthermore, the parameter space is discretised into bins, which introduces an error in understanding where the AV will pass/fail across the entire parameter space. Tu et al. [54] present a coverage-based testing approach that is effective for determining performance boundaries of ADS sub-systems, especially the planner. They perform this by estimating the probability of failure across the entire parameter space, acknowledging that the discretisation becomes inefficient in higher dimensions when no further measures are taken. Another coverage-based testing approach is presented by Li et al. [58,59], who use a limited set of scenarios to evaluate the performance of an AV, coined as the few-shot testing problem. They formulate this as an estimation problem and optimise the evaluation error by using a surrogate model instead of actual ADS. Overall, this highlights the importance of validation strategies considering a realistic and finite testing budget (both in terms of cost and time). However, using an SM instead of an actual ADS to optimise the scenario set poses limitations for more complex ODDs and respective ADS implementations.

Since the scenario parameter space is infinite due to the CPs of the LS, every coverage-based testing method needs to address parameter discretisation to arrive at CSs to test. Mori et al. [60] showcase that particular discretisation processes cannot provide sufficient argumentations for safety validation. However, the underlying experimental setup focuses purely on perception. In the discretisation process, one can utilise the respective parameter distributions of the CPs for each LS. These parameter distributions can, for example, include the likelihood of exposure to given scenarios. As opposed to using uniform distributions, realistic parameter distributions enable a more efficient design of the sampling process. For example, they consider that higher probabilities of exposure in the real world are highly significant for the safety validation of ADSs. Kaiser et al. [61] state the possibility of dividing the probability density function (PDF) of a CP into equivalence classes, either in an equidistant manner or using other methods. Furthermore, there are a few more methods to increase efficiency for these assumptions (see [62], or [63]). In [63], different concepts for the coverage calculation are based on the past scenario in the overall scenario space. A method for drawing efficient test case samples, independent of ADS implementations, will be shown in Section 4.3 based on [23].

A coverage-based testing method requires a dedicated process to determine the ODD coverage. In [39], such an ODD coverage process is proposed. Since it is essential to understand the respective assumptions and limitations when determining the ODD coverage to correctly utilise this information for the risk acceptance criteria proposed in Section 5, Section 4.1 briefly explains this process. Figure 3 displays the most basic idea of the concept. Theoretically, one can determine the coverage as the ratio of passed and failed executed scenarios [23].

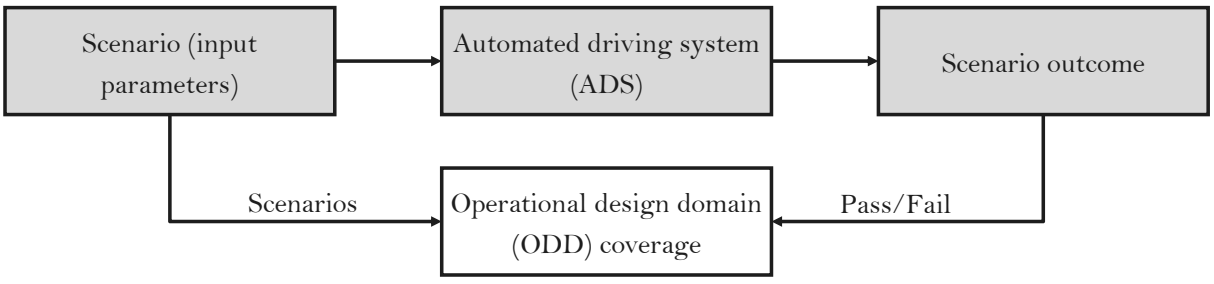

Figure 3. The achieved ODD coverage depends on the ratio of passed and failed executed scenarios [23].

A coverage-based testing method should be independent of specific ADS implementations, including different versions and manufacturers. Furthermore, it should be independent of the test method, and, therefore, it is possible to use more than one specific test method, not only simulation, which opens up possibilities regarding the allocation of scenarios to test methods. Also, the extension towards various target ODDs is required. In addition, a coverage-based testing method should provide the necessary values for the decision of an AV deployment in an interpretable and explainable manner. Hence, concrete criteria are required and will be presented in Section 5 as part of this work. Finally, the underlying process should cover the target ODD of an ADS in an efficient and traceable manner. Section 4.1 gives more details regarding such a process for coverage evaluation.

4.1. Exploring the Process of Coverage Evaluation

As mentioned in the previous section, Ref. [39] presents an ODD coverage process. Only if the coverage-based testing strategy can practically provide the necessary ODD coverage aspects can determining the risk acceptance criteria in Section 5 be viable. At first, the target ODD is specified, describing the operating capabilities of the given ADS, using a top-level taxonomy such as [4]. The next step introduces the concept of disturbances as defined in [32], enhancing the SBT approach by acknowledging the underlying physical principles of ADS operation. Essentially, the disturbances defined in the previous step are used to generate the respective LS, and using available data, the individual CPs are defined using PDFs.

This approach indirectly allows different traffic periods to be modelled into the scenario design, e.g., by adapting the scenario parameter PDFs. In addition, the current traffic flow could be considered an ODD parameter, setting certain operating limits in the first step of the process. As the next step, concrete scenarios are generated using a dedicated strategy for the discretisation process as previously explained and further detailed in Section 4.3. Using such a discretisation process also allows the inclusion of parameter restrictions (e.g., certain speed restrictions) for a single parameter. However, this is currently not possible for multiple parameters or even certain combinations and needs to be expanded on in the future. Overall, such parameter restrictions can be considered at the ODD, BC, or scenario level.

The next step evaluates the concrete scenario based on distinct, mostly safety-related, metrics. The last step evaluates the actual achieved ODD coverage, leading to specific follow-up actions, such as immediate ADS deployment, adoption of the target ODD, or adjustment of the testing effort.

4.2. Defining Coverage in the Evaluation Process

Depending on the concrete domain, there are many different definitions for the term coverage. For example, in software, the term coverage is used for testing purposes or for determining the covered amount of source code executed with a specific test suite [15,64]. Furthermore, in dependable and secure computing, the following definition for coverage is introduced by [65]:

> *Coverage refers here to a measure of the representativeness of the situations to which the system is subjected during its analysis compared to the actual situations that the system will be confronted with during its operational life.*

Such a definition of coverage is already very close to interpretations of coverage used in the ADS domain, although it is still general and unspecific. In the ADS domain, different coverage interpretations exist. Zhang et al. [53] discuss a selection of possible coverage interpretations. For example, one potential way to interpret coverage is to connect it to exploration within the given scenario space, while another is to define it in relation to the critical scenarios covered. These two interpretations are quite different in terms of the overall goal and the required methods to maximise either one or the other coverage. These interpretations also showcase the difficulty of comparing methods and implementations that use different definitions for coverage.

Overall, in the ADS domain, most coverage definitions are based on simple pass/fail ratios, which insufficiently cover the coverage aspect. A proper definition of coverage needs to be based on the ODD since this is the overall aim, based on the safety argumentation, and therefore, the definition needs to be suitable for that. Hence, a domain-specific definition of coverage needs to be embedded into an overall SBT-driven process for generating evidence for the safety argumentation. Such a definition is established in [23,30]. It aligns with the SBT approach for ADS safety validation, as it is based upon LS descriptions defined with PDFs (based on specific metrics) for the respective CPs. The following section briefly explains this.

As mentioned, the following coverage definition aligns with an SBT approach and is based on [23,30]. Based on this notion, the defined target ODD and the ADS behaviour competencies lead to relevant LS. These LSs are defined using CPs with PDFs based on certain metrics, such as the occurrence probability based on real-world data. Since the CPs of each LS define a parameter space, a respective discretisation needs to decide on the exact placement of the test case values to define the CSs, which are subsequently used to test the AV. On the CP level, for each test case value (marked with red dots in Figure 4), the individual coverage contribution is defined by the following equations, which is the area under the PDF between two distinct values $x_{i,a}$ and $x_{i,b}$, defined by the parameter discretisation process:

$$Pr(x_{i,a} \leq X \leq x_{i,b}) = \int_{x_{i,a}}^{x_{i,b}} f_X(x)dx. \tag{1}$$

These values are used to determine the coverage contribution of each CS, which are then aggregated to the LS and ODD levels, as shown in Figure 5a,b. The detailed equations are given in [23]. Furthermore, Figure 5c showcases how, after sorting all CSs based on their individual coverage contribution, a cumulative distribution function (CDF) can be constructed. Based on this, a PDF can be generated. This approach will be essential to the risk acceptance criteria in Section 5.

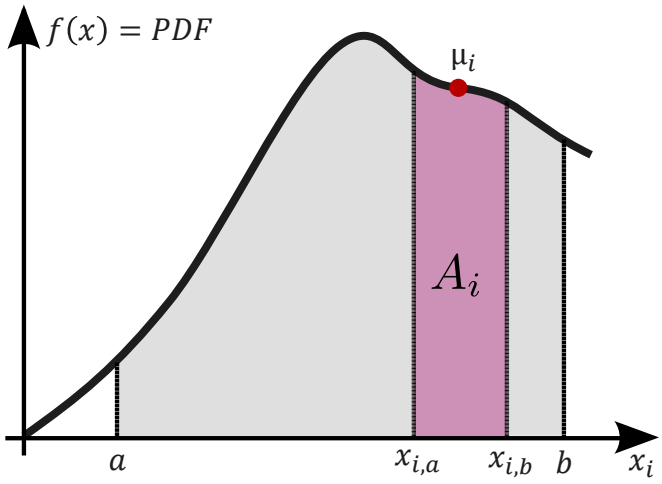

Figure 4. Example PDF of a continuous parameter x as part of a concrete scenario. The area A_i equals the area under the curve between x_i, a and x_i, b and defines the coverage for that parameter range [23].

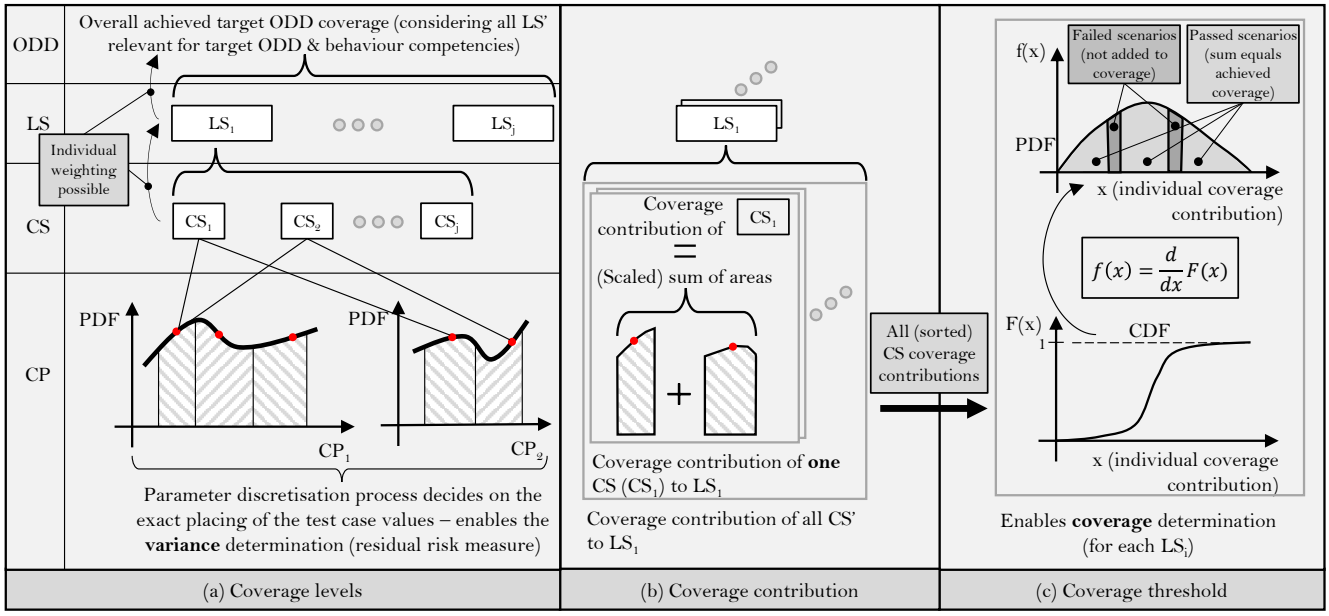

Figure 5. Overview of the concept for defining the target ODD, including the parameter discretisation, based on [23]. (**a**) The different coverage levels, starting from the CP level, where the parameter discretisation occurs (red dots represent the chosen test case values), up to the CS, LS, and the overall target ODD. (**b**) The individual coverage contribution of two CPs towards one specific CS, which is part of an LS. (**c**) Here, the sorted CS coverage contributions can be combined into a CDF and PDF, which defines the respective thresholds.

4.3. Proposed Sampling Method for Efficient Test Case Generation

As presented in the previous chapters, an SBT-driven coverage-based testing method (e.g., [39]) requires a parameter discretisation process providing a traceable approach towards generating CS from given LS descriptions. The following sampling method is in line with the provided definition of coverage in Section 4.2. It enables the determination of the covered areas of LS', and thus, coverage of the target ODD. The sampling method is presented in [23] and briefly explained as a prerequisite for the risk acceptance criteria presented in Section 5.

In line with the coverage definition of Section 4.2, the presented sampling method uses prior information at the CP level to determine the respective test case values (red points in Figure 5). Concretely, this prior information can come from sources like real-world data, actual AV operation, or virtual testing. Utilising prior information reduces the necessary testing (quantified using the proposed sampling method) while simultaneously enabling parallel execution of the CSs with no restriction regarding the test methods, which also allows efficient test case allocation. In comparison, most other methods solely rely on virtual testing to perform optimisation-driven coverage-based testing, such as determining performance bounds. This is problematic because current virtual testing frameworks lack the necessary fidelity to accurately represent the complete target ODD (e.g., regarding realistic environmental effects). Thus, the gap between the simulation and the real world can falsify the outcome and eventually guide the optimisation in incorrect ways. Hence, the outcome of such simulation studies should only be used to feed the prior CPs of the respective LS to reduce the test effort rather than to determine the final deployment decision.

The following presents the basic principle of the sampling method for the one-dimensional case. This means that the PDF is solely defined for one specific CP, assuming all other CPs of one particular LS are independent of this CP. More details, including the respective algorithms and benchmarks against other sampling methods, can be found

in [23]. Essentially, such a sampling method needs to answer the following question: how can the parameter range be discretised to obtain concrete values needed for CS definition and generation?

To enable an overall calculation of the achieved ODD coverage, each concrete value is assigned to a respective area under the PDF—or put otherwise, for a respective PDF area, one representative value, most of the time this will be the expected value (mean, assuming a random variable), will be chosen. Therefore, the aim should be to choose values that are as follows:

- As representative of the respective PDF area as possible. Hence, the variance needs to be reduced.
- As "different" from the other chosen values as possible. Hence, the in-between-variance needs to be maximised.

Overall, this leads to choosing the unsupervised learning technique k-means clustering to generate such values (the centroids of the clusters). Considering the relationship between the different variances leads to the following equation:

$$V_{total} = W + B, \tag{2}$$

where V_{total} is the overall variance in the PDF, W is the mean of the variance in each cluster (within-variance), and B is the variance in the centroids (between-variance). The clustering maximises B (by minimising W, as V_{total} stays the same). This also showcases the importance of using a prior, as V_{total} is reduced (and therefore W).

To summarise, the sampling method aims to achieve two goals. The first one is to minimise the distance between the occurring and chosen values. This can be reformulated to an unsupervised learning problem, where k-means is chosen. This is because k-means is a well-known method with many optimised implementations available, and it scales well towards higher dimensions. This allows scaling towards more complex scenarios, with more parameters influencing each other.

Figure 6 shows this in the first two steps. The second goal concerns individual test cases where the initial k-means clustering solution is adapted to stay within the pre-defined variance boundaries. Figure 6 shows this in the third and fourth steps. As mentioned in the beginning of this section, further details can be found in [23].

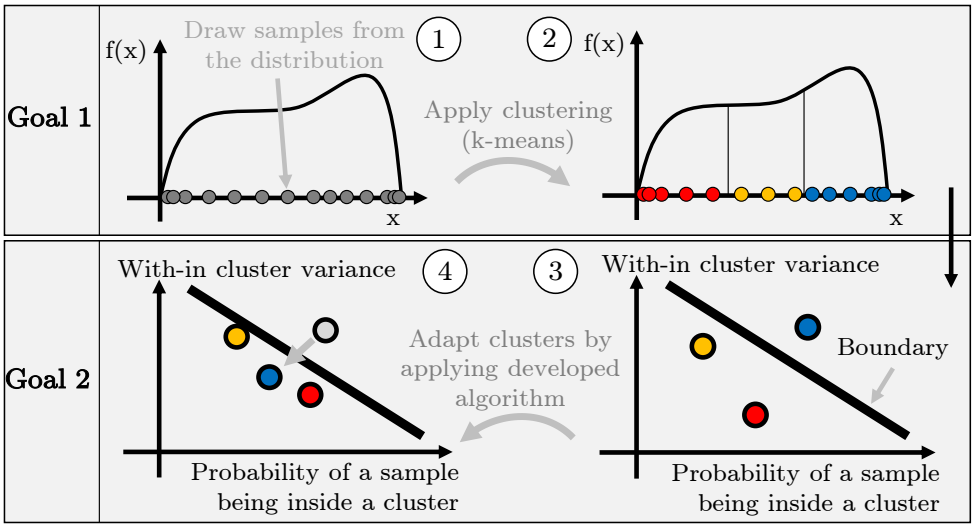

Figure 6. The proposed method for efficient sampling from PDFs defined for CPs of LS descriptions. In the first two steps, samples are drawn from the distribution, and k-means clustering is applied. In the third and fourth steps, the initial clusters are adapted to comply with the defined boundary condition. Illustrated is the one-dimensional case. The detailed algorithm is part of [23].

5. Risk Acceptance Criteria for the Safety Argumentation

In theory, after applying all relevant safety measures along the development process of an ADS—including the necessary safety factors—to determine the validation stop criteria, there will be a remaining residual risk [40]. The industry and, ultimately, society need to answer the question of what an acceptable level of residual risk is [46]. However, providing new risk acceptance criteria (RACs) can help answer this question, especially regarding the actual AV deployment decision and the complete ADS life cycle. Also, in the context of explainable AI for automated driving, Ref. [66] states that quantitative acceptance criteria, as requested by the SOTIF standard, still need to be included. This further amplifies the motivation for explainable risk acceptance criteria in the safety argumentation.

In Section 3, the need for RACs to decide on AV deployments is already explained. Furthermore, Section 4 presented a coverage-based testing method, including a sampling strategy for the parameter discretisation—both based on past work [23,39]. These methods provide the necessary foundation to define meaningful RACs for use in the pre-deployment phase of AV development and to make informed decisions about potential AV deployments.

In principle, residual risks are introduced inherently into the overall safety validation process. Quantitizing this introduces uncertainty as a residual risk, and one can control it accordingly. Looking at the ODD coverage process reveals residual risks throughout the process. For example, the residual risk between the TOD and the ODD (and thus the potential risk for an ODD extension) is tackled in [67].

The proposed RAC in this work covers a particular residual risk for coverage-based testing methods as part of an overall ODD coverage process. Since each relevant LS is defined by a parameter space, an infinite amount of CSs is possible. Hence, parameter discretisation needs to be performed, which introduces a residual risk of missing certain parameter combinations, which could reveal unsafe AV behaviour. The proposed RAC can guide and help determine if and how the discretisation affects the residual risk. Such quantification becomes possible using a coverage-based testing safety assessment strategy, with a respective parameter discretisation allowing such estimates. Concretely, the bounded variance in the parameter discretisation process, which is included by design in the sampling method, enables a bounded variance for each CS (see Section 4.3 or [23] for details). Defining the required levels for coverage and the boundaries for the variance enables the determination of the required amount of CS. These coverage and variance levels can be defined for each LS separately. It needs to be explicitly mentioned that this RAC does not represent one value (e.g., the performance of an AV or the overall accident probability) but gives an overall perspective of the depth of testing of an LS and, hence, the target ODD. Hence, the RAC contains two proposals for acceptance criteria, which can also be combined. Eventually, this will provide an explainable way to argue for the achieved coverage of the target ODD relevant to the potential AV deployment.

5.1. Acceptance Criteria

As Section 4 has shown, it is necessary to define respective methods to derive meaningful acceptance levels for coverage. Coverage and variance together form a meaningful macroscopic assessment, which can be used to compare different ADS solutions or versions, including changes in behaviour competencies or the target ODD. Hence, two different criteria are presented, one for the coverage and one for the variance. Both strongly rely on a coverage-based testing method for ADS safety validation, including a respective parameter discretisation process. The resulting residual risk can be quantified using the two criteria presented next.

5.1.1. Coverage Criteria

Having defined the term coverage across all required levels (see Section 4.2, Figure 5, and [23] for the details) enables to set a dedicated threshold for the required coverage. However, this begs the question: how do we define and argue for such a threshold? In a mathematical expression, this looks like the following equation:

$$Coverage_{achieved} \geq Coverage_{threshold}, \tag{3}$$

where $Coverage_{achieved}$ stands for the actual achieved coverage, based on the testing results, and $Coverage_{threshold}$ stands for the respective coverage threshold. Being able to calculate the individual coverage contribution for each CS of an LS enables one to compute the respective (discrete) CDF (see Figure 5c). This, in turn, enables the determination of the PDF. Both shapes of CDF and PDF are determined by the underlying CPs PDFs, and will, therefore, differ for each LS. The coverage level could be defined as the percentage of the area covered under the PDF. Hence, assuming a Gaussian distribution for the PDF, a concrete and meaningful threshold would be the x-sigma levels. For example, the threshold for μ ± 3 sigma would be 99.73%. Even if the actual PDF is not Gaussian, the value (e.g., 99.73% in the 3-sigma level case) is still useable and meaningful, as it means that 99.73% of the PDF area is covered.

5.1.2. Variance Criteria

Having defined the term coverage across all required levels and using a dedicated sampling method (see Section 4.3, Figure 6, and [23] for the details) enables to set a maximum threshold for the variance of each resulting CS. However, how do we define—and argue for—such a threshold? Once again, the mathematical expression can be formulated as

$$Var_{CS,i} \leq Var_{threshold} \tag{4}$$

where $Var_{CS,i}$ is the resulting variance of the i-th CS, and $Var_{threshold}$ stands for the respective variance threshold. The resulting CS PDF is different for each CS and depends on each CP and its individual PDF, which is part of the respective LS. Figure 5b shows this for one specific CS qualitatively. Such a PDF can be determined a posteriori (after the parameter discretisation) for each CS but is not known prior. Hence, thresholds for the variance without knowledge of the PDF are difficult. Overall, the variance is considered a weak risk measure since it cannot accurately summarise all risk aspects in all relevant circumstances [68]. However, it can be compared against other variance values, as smaller variances are preferred. This can be explained by the fact that in the case of smaller variances for a certain parameter section that one specific test case value should represent, there is a clearer notion of which area is to be considered more important since this depends on the shape of the PDF of the individual CP, which is based on the prior information. Since, due to the necessary parameter discretisation, one representative test case value has to be chosen for each parameter section, smaller variances are preferred.

For each LS, n CPs are defined by their probability distribution. Therefore, the following assumption (for each CS) is made:

$$X_1 + X_2 + \cdots + X_n = Y_{CS,i}, \tag{5}$$

with X_i being the random variable with the associated probability distribution from CP_i and Y being the sum of all CP random variables. Since it is defined that the individual CPs are independent (for now, as only 1D is considered; however, an extension to ND is possible as well), the following can be stated:

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i}^{n} Var(X_i). \tag{6}$$

From this equation, it follows that the overall variance in CS_i is bounded if the variance in each test case value $X_{1,\dots,n}$ for the individual $CP_{1,\dots,n}$ is bounded as well. The overall variance is the result of a linear combination:

$$Var(X_1) + Var(X_2) + \cdots + Var(X_i) = Var(Y). \tag{7}$$

If the variances in the individual X_i can be described as a function of the occurrence probability $p_{c,i}$, the overall variance in each CS becomes a function of the probability of occurrence.

Therefore, the definition of the variance boundary should be on the CP level for each CP. As a baseline, a well-known distribution that assumes no prior information can be used. Concretely, this could be a uniform distribution with range [0, 1], $\mathcal{U}_{[0,1]}$, as it has a well-defined variance. Using the range [0, 1] can be argued without restriction of general validity, as, for example, the sampling technique in [23] resamples and scales everything to the [0, 1] range. In doing this, the achieved variance level can be compared with the number of required test cases to achieve the same variance for $\mathcal{U}_{[0,1]}$. This enables the showcase of the test case reduction potential of using a prior and presents itself as an interpretable way to compare the achieved variance levels.

To achieve meaningful results, all test cases' overall variance must be weighted based on their occurrence value (which represents the area under the PDF; see Figure 4). For $\mathcal{U}$, everything stays the same, as all generated test cases have the same amount of occurrence probability. For the CP with a PDF that is different from $\mathcal{U}$ (and therefore uses a prior), this acknowledges that values with higher occurrence are more important (defined by the prior). Concretely, this means

$$W_{c,i,weighted} = \sum p_{c,i} W_{c,i}, \sum p_{c,i} = 1, \tag{8}$$

with $W_{c,i,weighted}$ being the weighted within-variance, $W_{c,i}$ being the within-variance, and $p_{c,i}$ being the occurrence probability of one specific test case value. This also leads to the desired behaviour of CSs with higher coverage contributions and lower variance levels. Otherwise, comparing the uniform distribution is unfair and meaningless when only the maximum variance value is used. A certain maximum value for the allowed variance can be set nonetheless, e.g., using a line as the boundary condition. This is already implemented in the sampling method presented in [23]. The boundary line defined in the sampling method considers this by forcing lower variance values to test case values with higher occurrence probabilities. Therefore, it minimises this weighted variance to a certain degree.

A comparison of the achieved variance levels with the (fractioned) $\mathcal{U}_{[0,1]}$ can be used to showcase the effectiveness of the used sampling method or, in general, the depth of the executed testing effort. In principle, one can showcase that fewer than j test cases are required to achieve the same variance levels as the $\mathcal{U}_{[1/j]}$ (the fractioned uniform distribution, with each fractions covering the parameter range of $1/j$). Defining the risk measure based on the achieved variance levels compared to the $\mathcal{U}_{[1/j]}$ leads to the risk measure being a function of j, with j being the number of test cases for the $\mathcal{U}$ case. Additionally, this represents the residual risk, as higher j means lower variance, which is subsequently a lower risk measure and a lower residual risk. Eventually, this enables a combined view of coverage level and risk measures.

Figure 7 qualitatively compares the resulting within-variance of $\mathcal{U}_{[0,1]}$ with a different number of sections (representing j) with exemplarily sampled test cases using the sampling method from [23] (see also Section 4.3). Each big black dot represents one test case value, with the individual occurrence probability displayed on the x-axis. The y-axis represents the within-variance. The boundary line mentioned above is also included. Overall, this showcases the different levels of within-variance for different $\mathcal{U}_{[1/j]}$ compared to another sampling technique using prior information.

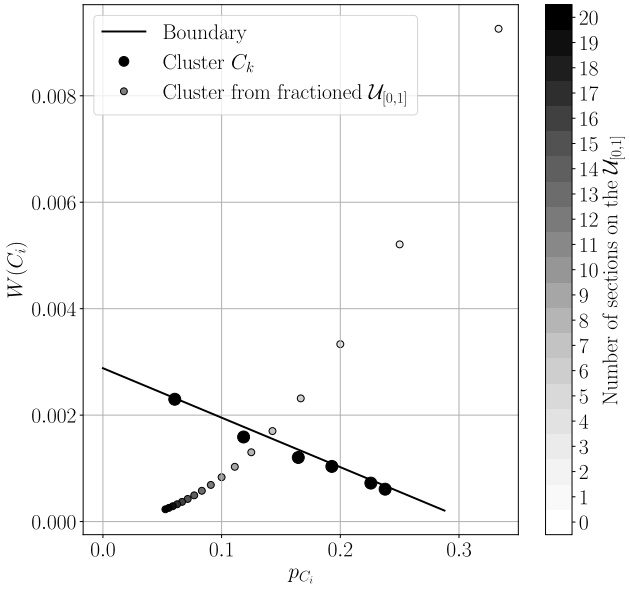

Figure 7. A qualitative comparison between the within-variance of the fractioned $\mathcal{U}_{[0,1]}$, with the number of sections ranging from 0 to 20, and a result of the sampling technique from [23] is displayed.

5.1.3. Combining Coverage and Variance Criteria

The overall safety argumentation regarding AV safety in the target ODD is a combined argument based on the achieved ODD coverage for a particular risk measure. Hence, the additional risk measure information can distinguish certain ODD coverage values (or compare certain coverage values). As a simple example to showcase the necessity of this, one can imagine the following situation, assuming the same target ODD. For one ADS implementation, its provider claims to have achieved total coverage of the target ODD. However, this may only be achieved by not testing accurately enough, leaving much room for uncertainty. Another ADS provider may also claim to have achieved nearly total coverage for another ADS implementation. However, because this ADS implementation is tested much more accurately, the uncertainty of error is reduced. Hence, only introducing the additional risk measure based on variance allows for a meaningful comparison of ADS implementations using coverage-based testing.

Furthermore, this is associated with a deterministically determined amount of required test cases to achieve certain risk measure levels (and coverage levels; however, this also depends on the respective ADS performance and not solely on the test effort) assuming $\mathcal{U}_{[0,1]}$ for all underlying CPs. Hence, an explainable way to achieve the necessary number of test cases is achieved. As previously explained, by using prior information, a reduction in test cases can be achieved, and by comparison with $\mathcal{U}_{[0,1]}$, quantifiable and explainable comparison is enabled. Additionally, a reasonably well-defined boundary separating the acceptable and non-acceptable areas (in terms of the safety argument) can be established. A simple example of such a function could be

$$f(r) = 1 - \frac{1}{250x + 10}, r = \frac{1}{j}, \tag{9}$$

where x is the residual risk, and the resulting $f(r)$ is the ODD coverage. The achieved safety margin can also be quantified as the perpendicular distance to the separating function. The following section (Section 6) further showcases this with an application example.

6. Application Example

A simplified example consisting of two LSs is presented below. This application example aims to showcase how scenario design can influence the prior. It thus leads to test case reduction while simultaneously providing an explainable argument for the ODD coverage required for AV deployment. The example is based on [30]. Based on the presented scenario design and overall setup, the RACs of Section 5 can be applied to quantify the test case reduction in a proper manner. In principle, two LSs are defined at a specific intersection in the INTERACTION dataset [69]. For each LS, the traffic participant's starting position is chosen. Based on this, one distinct lanelet (a data format to represent road networks, see [70]), which is relevant for the future trajectory of the traffic participant, is chosen. The available data in the INTERACTION dataset from all recorded traffic participants are analysed to compile PDFs for the velocity, heading, lateral position, and time offset (which is used to delay the start of the traffic participant in the scenario). These four CPs define the individual LS. Figure 8 displays the first LS (LS 1) and the second LS (LS 2).

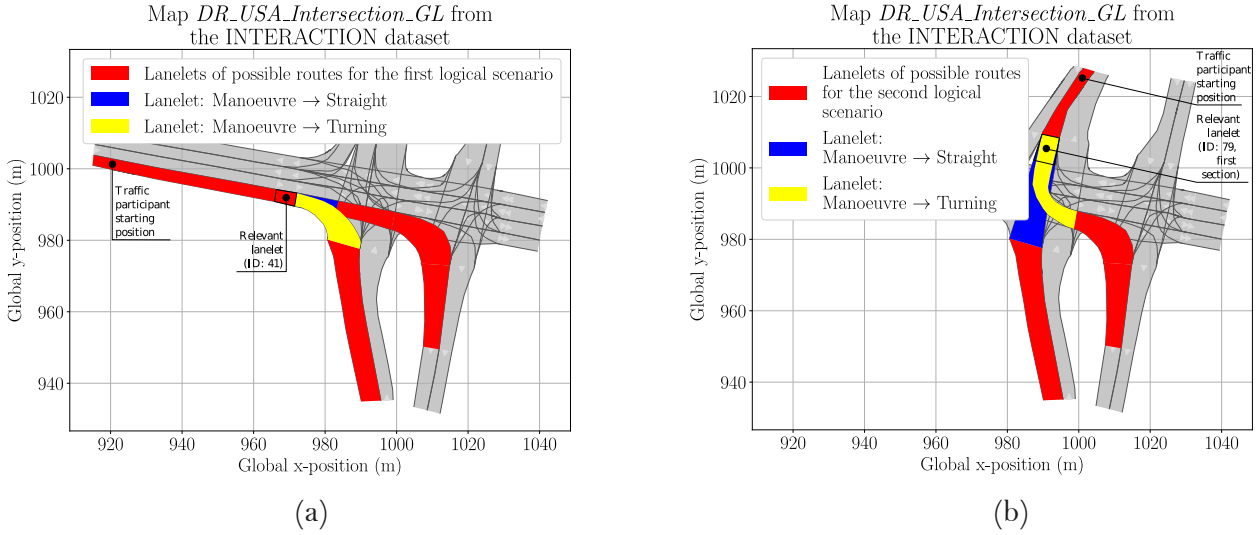

Figure 8. (**a**) The starting position of the traffic participant and the considered routes for the LS 1. The extracted data from the relevant lanelet with ID: 41 are used to construct the PDFs for LS 1. (**b**) The starting position of the traffic participant and the considered routes for LS 2. The extracted data from the relevant lanelet with ID: 79 are used to construct the PDFs for the LS 2.

The generated PDFs for the CPs are used as prior information. For LS 1, analysis of the data at the distinct lanelet with ID: 41 show a high correlation between lateral position and heading of the traffic participant. This is shown in Figure 9 and utilised to construct a 2D PDF to perform the sampling.

Using the sampling technique explained in Section 4.3 (which is presented in detail in [23]) enables the generation of test case values that are much more efficient compared to using no prior knowledge. This can be quantified using the RACs from Section 5. Overall, the performance of the sampling method is quantified using the weighted within-variance defined in Equation (8) and compared with the uniform distribution case. This enables determining how many test cases are necessary, in the case of a uniform distribution (no prior), to reach the same level of weighted within variance. Hence, a meaningful approach to quantifying the test case reduction potential of utilising prior information on the CP level is achieved. Concretely, Table 1 shows the detailed numbers. LS 1, which uses the 2D prior, achieves a reduction of 83.23%, while LS 2 achieves 75.89%. Each row shows the exact achieved numbers for each LS—with and without a prior. LS 1 achieves this reduction mostly because a 2D prior for lateral position and heading is utilised, whereas LS 2 achieves the reduction due to a combination of individual reductions for each CP.

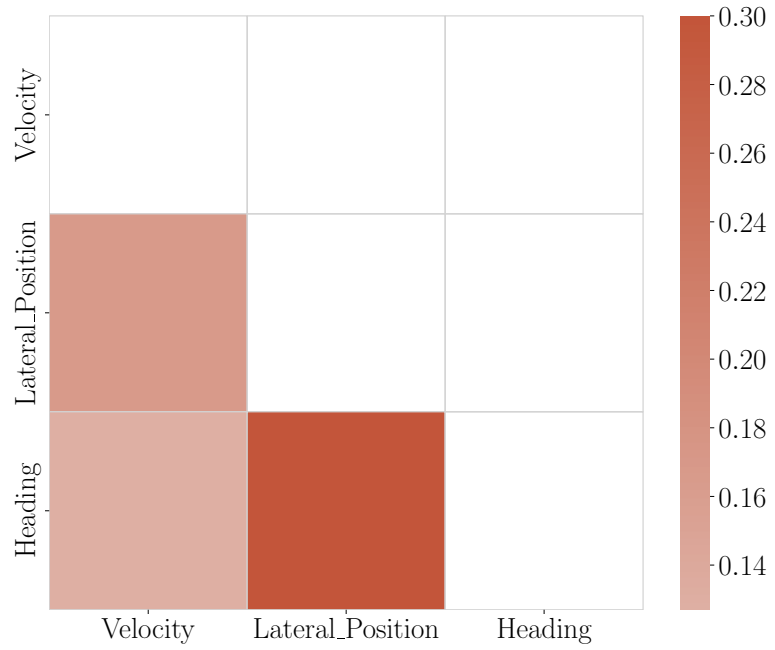

Figure 9. The correlation between the individual features extracted from the INTERACTION dataset (for lanelet with ID: 41) is displayed. An exceptionally high correlation between the lateral position of a vehicle (in the respective lanelet) and its heading angle can be observed and utilised.

Table 1 shows that LS 2 contains four individual CPs. For the case of no prior knowledge, a uniform distribution for each CP leads to a high number of required test cases to achieve certain thresholds for the residual risk. This is shown in Figure 10 concretely, as different values for the residual risk and the resulting amount of test cases are displayed. The values for the residual risk at the x-axis of Figure 10 are exemplary. Based on these values and the assumption that LS 2 contains four individual CPs, the variance is a fourth of the respective residual risk value. Assuming a uniform distribution allows us to calculate the resulting test cases in Figure 10.

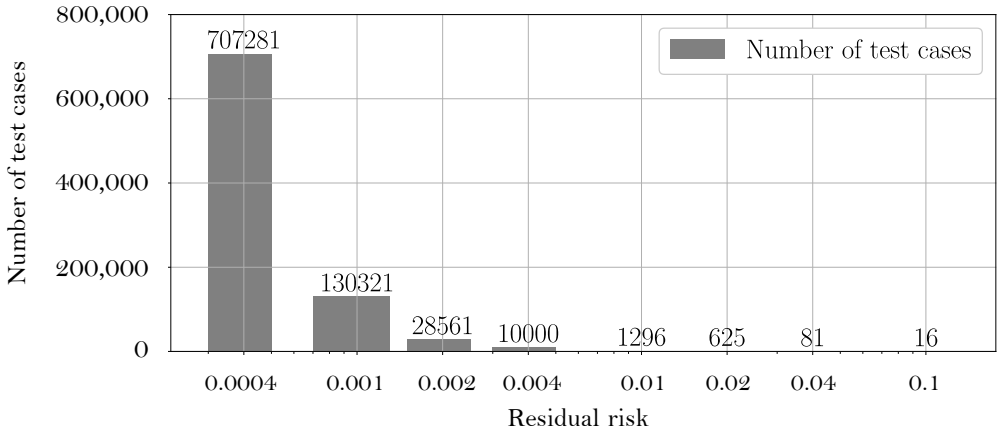

Figure 10. The required amount of test cases to achieve a certain residual risk assuming four individual CP with a uniform distribution.

Table 1. Both considered LSs and their respective properties, the number of test cases, and the test case reduction.

Parameter	Dimensions	LS 1 (With Prior)	LS 1 (Without Prior)	LS 2 (With Prior)	LS 2 (Without Prior)
Velocity	1D	6 values	7 values	6 values	8 values
Heading	1D	-	-	6 values	8 values
Time offset	1D	3 values	4 values	3 values	6 values
Lateral position	1D	-	-	6 values	7 values
Lateral position & heading	2D	6 values	23 values		-
Number of test cases	-	108	644	648	2688
Test case reduction	-	83.23%	-	75.89%	-

7. Discussion

The presented example shows two main aspects: Firstly, using prior information to construct the overall LS, concretely as PDF on the CP level, offers great reduction potential in terms of the required number of test cases. Secondly, quantifying such a test case reduction becomes possible and meaningful with the RACs of Section 5, concretely, the variance criteria implemented as the weighted within-variance. In addition to these two aspects, combining the two acceptance criteria introduced in Section 5 offers further potential. This is shown in Figure 11 in a qualitative example. Using the achieved residual risk values of both the LS and a simple equation to define a separating function enables distinguishing between acceptable and non-acceptable areas in terms of residual risk and coverage combinations. Concretely, the necessary threshold for coverage can be determined for given values of residual risks and vice versa. For example, for a residual risk of 0.02, a coverage level of at least 93.3% is required (based on Equation (9)). In addition, a safety margin for the coverage (but in principle also for the residual risks) can be defined. The separating function is chosen to require total coverage in cases of very high residual risk and vice versa, with a minimum amount of coverage required regardless of residual risk. This function is defined in a simplified manner and should showcase the possibilities of using the acceptance criteria, coverage, and variance together.

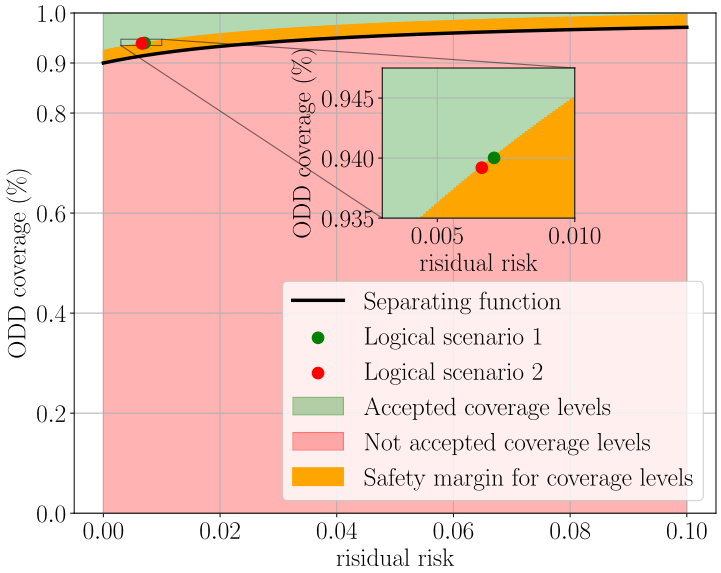

Figure 11. The combination of residual risk and ODD coverage (simplified based on LS coverage) displayed for the presented LS 1 and 2, including the separating function. Based on the achieved residual risk, threshold values for the coverage can be determined (and vice versa).

8. Conclusions and Outlook

AVs are on the brink of introduction, with deployments in specific ODDs already happening across the globe. However, the safety validation of ADS-equipped vehicles is still a significant issue. Among the different existing validation strategies, coverage-based testing offers a possibility of determining ADS safety across the target ODD. Current methods mainly rely on pure simulation studies to determine performance boundaries instead of providing methods which do not rely on specific assumptions regarding the test method and offer a way to determine ADS safety.

This work showcases that meaningful RACs can be defined for safety argumentation as part of the deployment decision. The RACs directly lead to a more explainable and interpretable way to argue for ADS safety. This is based on an existing coverage-based testing method that utilised a sampling technique to generate CS with bounded variance by design. The notion that a CS with a greater coverage contribution should have a lower variance can be included on the CP level using the boundary line of the applied sampling method. Hence, quantifying such a test case reduction becomes possible and meaningful with the proposed RACs. Furthermore, using prior information enables a real test effort reduction potential. In addition, the definition of threshold levels enables the determination of the CS amount and can act as a viable risk measure.

However, the application possibilities go beyond one specific ADS. Concretely, using the proposed RACs enables a comparison between different sensor setups or ADS versions and, in principle, also between completely different ADS. Regarding the overall validation process, the proposed RACs can act as validation stop criteria due to the enabled quantification of coverage and residual risk—explicitly regarding the parameter discretisation. This provides meaningful measures to quantify necessary test efforts for specific target ODDs and potential ODD extensions. Another vital aspect to mention is scalability: every proposed method for ADS safety validation needs to scale to massive amounts of LS' and respective CPs. The proposed RACs in this work also provide meaningful measures for these cases, especially considering the use of prior information.

Multiple future research directions are possible based on this work's findings. This work has shown that using prior information offers great test reduction potential. Focusing on the prior information for the sampling strategy to further increase efficiency, e.g., using methods to determine reasonably foreseeable parameter ranges, can combine different coverage-based testing methods to achieve ADS safety validation. Also, cases where real data are scarce need to be considered. This negatively affects the prior knowledge and potentially leads to reduced test case reduction potential. Various aspects to mitigate such situations need to be explored in future research. In addition, it needs to be mentioned that further exploration is required to accurately ensure that the incorporated knowledge fulfils the required standards in terms of quality and relevance so as not to guide the testing effort in the wrong direction. Hence, only managing the prior well guarantees meaningful test case reductions.

Furthermore, the RACs presented in this work can benchmark the sampling strategy against other methods (e.g., performance boundary estimation) in light of potential insufficiencies across the entire parameter space, not only due to physical boundaries. In addition, by extending the approach for combining the achieved coverage of different LS towards the target ODD, including the behaviour competencies, a joint coverage of the target ODD and behaviour capabilities of ADS-equipped vehicles can be achieved. Additionally, the residual risks need to be quantified not only for the parameter discretisation process but throughout the ADS life cycle. This must include dedicated methods for estimating and handling pre- and post-deployment risks. In line with this, respective RAC thresholds across the different ADS life cycle phases need to be defined. While this article proposes a concrete RAC for the parameter discretisation, this must be further discussed and explored in future research.

Eventually, this needs to lead to a robust, explainable, scalable, and updatable continuous safety argumentation across the life cycle by further investigating failure rates, including the respective acceptance criteria. Since ADSs include AI models, this proposed validation strategy can be exploited for AI systems across multiple domains for the respective assurance argument required for deployment.

Author Contributions: Conceptualisation, P.W. and G.S.; methodology, P.W. and G.S.; software, P.W.; validation, P.W.; formal analysis, P.W.; investigation, P.W.; resources, P.W. and G.S.; data curation, P.W.; writing—original draft preparation, P.W.; writing—review and editing, P.W. and G.S.; visualisation, P.W.; supervision, G.S.; project administration, G.S.; funding acquisition, G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Union's Horizon Europe Research and Innovation Program under Grant 101076754—AIthena Project.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Authors Patrick Weissensteiner and Georg Stettinger were employed by the company Infineon Technologies AG. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EU	European Union
ADAS	Advanced driver assistance systems
AV	Automated vehicle
ADS	Automated driving systems
TOD	Target operational domain
ODD	Operational design domain
SOTIF	Safety of the intended functionality
SBT	Scenario-based testing
CP	Continuous parameters
LS	Logical scenario
CS	Concrete scenario
UNECE	United nations economic commission for europe
NATM	New assessment/test method
AUR	Absence of unreasonable risk
GSN	Goal-structuring notation
ALARP	As low as reasonably practicable
GAMAB	Globalement au moins aussi bon
MEM	Minimum endogenous mortality
PRB	Positive risk balance
CDF	Cumulative distribution function
RAC	Risk acceptance criterion
PDF	Probability density function

References

1. World Health Organization. *Global Status Report on Road Safety 2018: Summary*; WHO: Geneva, Switzerland, 2018.
2. European Road Safety Observatory; European Commission. *Road Safety Thematic Report*; Technical report; European Commission-Directorate General for Transport: Brussels, Belgium, 2021.
3. *ISO/SAE PAS 22736:2021*; Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE International: Warrendale, PA, USA, 2021.
4. *ISO/AWI 34503*; Road Vehicles—Taxonomy for Operational Design Domain for Automated Driving Systems. International Organization for Standardization: Geneva, Switzerland, 2023.
5. Stettinger, G.; Weissensteiner, P.; Khastgir, S. Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. *IEEE Access* **2024**, *12*, 22718–22745. [CrossRef]

6. Automated Vehicle Safety Consortium. *AVSC Best Practice for Evaluation of Behavioral Competencies for Automated Driving System Dedicated Vehicles (ADS-DVs)*; Best Practice AVSC00008202111, SAE ITC; SAE International: Warrendale, PA, USA, 2021.
7. Poddey, A.; Brade, T.; Stellet, J.E.; Branz, W. On the validation of complex systems operating in open contexts. *arXiv* **2019**, arXiv:1902.10517.
8. Burton, S.; Hawkins, R. *Assuring the Safety of Highly Automated Driving: State-of-the-Art and Research Perspectives*; University of York: York, UK, 2020.
9. Greifenstein, M.; Güthner, H.; Scharfenberger, P.; Kauschke, P.; Herrmann, A.; Kuhnert, F. *The Evolution of Shared Autonomous Vehicles (SAV)*; PricewaterhouseCoopers GmbH: Frankfurt, Germany, 2024.
10. Draghi, M. *The Future of European Competitiveness: Part B-In-Depth Analysis and Recommendations*; Technical Report Part B; European Commission: Brussels, Belgium, 2024.
11. He, J.Y.; Cheng, Z.Q.; Li, C.; Xiang, W.; Chen, B.; Luo, B.; Geng, Y.; Xie, X. *DAMO-StreamNet: Optimizing Streaming Perception in Autonomous Driving*; IJCAI: California City, CA, USA, 2023; Volume 2, pp. 810–818, ISSN 1045-0823. [CrossRef]
12. SaFAD. *Safety First for Automated Driving*; Mercedes-Benz Group: Stuttgart, Germany, 2019.
13. Batsch, F.; Kanarachos, S.; Cheah, M.; Ponticelli, R.; Blundell, M. A taxonomy of validation strategies to ensure the safe operation of highly automated vehicles. *J. Intell. Transp. Syst.* **2020**, *26*, 14–33. [CrossRef]
14. Corso, A.; Moss, R.; Koren, M.; Lee, R.; Kochenderfer, M. A Survey of Algorithms for Black-Box Safety Validation of Cyber-Physical Systems. *J. Artif. Intell. Res.* **2021**, *72*, 377–428. [CrossRef]
15. *ISO 26262:2018*; Road Vehicles—Functional Safety. Technical Report. ISO: Geneva, Switzerland, 2018.
16. *ISO 21448:2022*; Road Vehicles-Safety of the Intended Functionality. Technical Report Edition 1, ISO/TC 22/SC 32 Electrical and Electronic Components and General System Aspects. ISO: Geneva, Switzerland, 2022.
17. Ulbrich, S.; Menzel, T.; Reschka, A.; Schuldt, F.; Maurer, M. Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; pp. 982–988. [CrossRef]
18. Menzel, T.; Bagschik, G.; Isensee, L.; Schomburg, A.; Maurer, M. From Functional to Logical Scenarios: Detailing a Keyword-Based Scenario Description for Execution in a Simulation Environment. *arXiv* **2019**, arXiv:1905.03989.
19. ASAM e.V. *ASAM Test Specification Study Group Report 2022*; Technical Report Version 1.0.0; ASAM: Hoehenkirchen, Germany, 2022.
20. Riedmaier, S.; Ponn, T.; Ludwig, D.; Schick, B.; Diermeyer, F. Survey on Scenario-Based Safety Assessment of Automated Vehicles. *IEEE Access* **2020**, *8*, 87456–87477. [CrossRef]
21. Brade, T.; Kramer, B.; Neurohr, C. *Paradigms in Scenario-Based Testing for Automated Driving*; ACM: New York, NY, USA, 2021; pp. 108–114. [CrossRef]
22. Neurohr, C.; Westhofen, L.; Henning, T.; de Graaff, T.; Möhlmann, E.; Böde, E. Fundamental Considerations around Scenario-Based Testing for Automated Driving. *arXiv* **2020**, arXiv:2005.04045.
23. Weissensteiner, P.; Stettinger, G.; Khastgir, S.; Watzenig, D. Operational Design Domain-Driven Coverage for the Safety Argumentation of Automated Vehicles. *IEEE Access* **2023**, *11*, 12263–12284. [CrossRef]
24. Birkemeyer, L.; King, C.; Schaefer, I. Is Scenario Generation Ready for SOTIF? A Systematic Literature Review. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; pp. 472–479. [CrossRef]
25. Bock, J.; Krajewski, R.; Moers, T.; Runde, S.; Vater, L.; Eckstein, L. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. *arXiv* **2019**, arXiv:1911.07602.
26. Bagschik, G.; Menzel, T.; Maurer, M. Ontology based Scene Creation for the Development of Automated Vehicles. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1813–1820. [CrossRef]
27. Allen, J.; Koo, W.; Murugesan, D.; Zagorski, C. *Testing Methods and Recommended Validation Strategies for Active Safety to Optimize Time and Cost Efficiency*; Issue: 2020-01-1348; SAE Technical Paper 2020-01-1348; SAE International: Warrendale, PA, USA, 2020; ISSN 0148-7191/2688-3627. [CrossRef]
28. Junietz, P. Microscopic and Macroscopic Risk Metrics for the Safety Validation of Automated Driving. Ph.D. Thesis, Technische Universität Darmstadt, Darmstadt, Germany, 2019. [CrossRef]
29. Junietz, P.; Steininger, U.; Winner, H. Macroscopic Safety Requirements for Highly Automated Driving. *Transp. Res. Rec. J. Transp. Res. Board* **2019**, *2673*, 1–10. [CrossRef]
30. Weissensteiner, P. Safety Argumentation for the Deployment of Automated Vehicles. Bachelor's Thesis, Technical University Graz, Graz, Austria, 2023. [CrossRef]
31. Galbas, R.; Nolte, M.; Eberle, U.; Hungar, H.; Mosebach, H.; Salem, N.F.; Schittenhelm, H.; Reich, J.; Kirschbaum, T.; Westhofen, L. *VV Methods Safety Assurance Position Paper*; Position Paper, Verification and Validation Methods; Bundesministerium für Wirtschaft und Klimaschutz: Berlin, Germany, 2024.
32. JAMA; SAKURA. *Automated Driving Safety Evaluation Framework Ver. 1.0-Guidelines for Safety Evaluation of Automated Driving Technology*; Technical Report; JAMA: Tokyo, Japan, 2022.
33. Wagner, N.; Weissensteiner, P.; Coget, J.B.; Eckstein, L.; Bracquemond, A. Common Methodology for Data-Driven Scenario-Based Safety Assurance in the HEADSTART Project. In Proceedings of the ITS European Congress, Lisbon, Portugal, 18–20 May 2020.

34. Ciuffo, B.; Mattas, K.; Galassi, M.C. *Safety Assurance of Automated Driving Systems-Raising the Level of Ambition*; European Commission-Joint Research Center: Brussels, Belgium, 2020.
35. Donà, R.; Ciuffo, B.; Tsakalidis, A.; Di Cesare, L.; Sollima, C.; Sangiorgi, M.; Galassi, M.C. Recent Advancements in Automated Vehicle Certification: How the Experience from the Nuclear Sector Contributed to Making Them a Reality. *Energies* **2022**, *15*, 7704. [CrossRef]
36. United Nations Economic Commission for Europe. *New Assessment/Test Method for Automated Driving (NATM)*; Submitted by the Working Party on Automated/Autonomous and Connected Vehicles ECE/TRANS/WP.29/2021/61; United Nations Economic Commission for Europe: Geneva, Switzerland, 2021.
37. Al-Turki, M.; Ratrout, N.T.; Rahman, S.M.; Reza, I. Impacts of Autonomous Vehicles on Traffic Flow Characteristics under Mixed Traffic Environment: Future Perspectives. *Sustainability* **2021**, *13*, 11052. [CrossRef]
38. Sinha, A.; Chand, S.; Wijayaratna, K.P.; Virdi, N.; Dixit, V. Comprehensive safety assessment in mixed fleets with connected and automated vehicles: A crash severity and rate evaluation of conventional vehicles. *Accid. Anal. Prev.* **2020**, *142*, 105567. [CrossRef] [PubMed]
39. Weissensteiner, P.; Stettinger, G.; Genser, S.; Watzenig, D. Operational Design Domain Coverage for the Safety Validation of Automated Driving Systems. In Proceedings of the Driving Simulation Proceedings, Strasbourg, France, 15–16 September 2022.
40. *UL 4600*; Standard for Evaluation of Autonomous Products. Standard for Safety; Underwriters Laboratories: Northbrook, IL, USA, 2022.
41. Kelly, T.; Weaver, R. *The Goal Structuring Notation–A Safety Argument Notation*; Citeseer: Princeton, NJ, USA, 2004.
42. Aurora. *Aurora's Safety Case Framework*; Aurora: Bay Area, CA, USA, 2023.
43. Schittenhelm, H. *VVM Safeguarding Automation–How to Ensure a Safe Operation of an Automated Driving System by a Methodological Approach?—An Interims Report*; Verification Validation Methods: Stuttgart, Germany, 2022.
44. *BS EN 50126:1999*; Railway Applications—The Specification and Demonstration of Reliability, Availability, Maintainability, and Safety (RAMS). BSI: London, UK, 1999.
45. Favaro, F. Exploring the Relationship Between "Positive Risk Balance" and "Absence of Unreasonable Risk". *arXiv* **2021**, arXiv:2110.10566. [CrossRef]
46. Kauffmann, N.; Fahrenkrog, F.; Drees, L.; Raisch, F. Positive Risk Balance: A Comprehensive Framework to Ensure Vehicle Safety. *Ethics Inf. Technol.* **2022**, *24*, 15. [CrossRef]
47. Di Fabio, U.; Broy, M.; Brüngger, R.; Eichhorn, U.; Grunwald, A.; Heckmann, D.; Hilgendorf, E.; Kagermann, H.; Losinger, A.; Lutz-Bachmann, M.; et al. *Ethic Commission: Automated and Connected Driving*; Technical Report, Report of Ethics Commission Appointed by the Federal Minister of Transport and Digital Infrastructure; Federal Minister of Transport and Digital Infrastructure: Berlin, Germany, 2017.
48. *ISO/TR 4804:2020*; Road Vehicles—Safety and Cybersecurity for Automated Driving Systems—Design, Verification and Validation. International Organization for Standardization: Geneva, Switzerland, 2020.
49. Favarò, F.; Fraade-Blanar, L.; Schnelle, S.; Victor, T.; Pena, M.; Engstrom, J.; Scanlon, J.; Kusano, K.; Smith, D. Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk. Technical Report. 2023. Available online: www.waymo.com/safety (accessed on 20 October 2024).
50. Blumenthal, M.S.; Fraade-Blanar, L.; Best, R.; Irwin, J.L. *Safe Enough: Approaches to Assessing Acceptable Safety for Automated Vehicles*; Technical Report; RAND Corporation: Santa Monica, CA, USA, 2020.
51. De Silva, D.; Alahakoon, D. An Artificial Intelligence Life Cycle: From Conception to Production. *Patterns* **2022**, *3*, 100489. [CrossRef]
52. Hawkins, R.; Picardi, C.; Donnell, L.; Ireland, M. Creating a Safety Assurance Case for a Machine Learned Satellite-Based Wildfire Detection and Alert System. *J. Intell. Robot. Syst.* **2023**, *108*, 47. [CrossRef]
53. Zhang, X.; Tao, J.; Tan, K.; Törngren, M.; Sánchez, J.M.G.; Ramli, M.R.; Tao, X.; Gyllenhammar, M.; Wotawa, F.; Mohan, N.; et al. Finding Critical Scenarios for Automated Driving Systems: A Systematic Literature Review. *arXiv* **2021**, arXiv:2110.08664.
54. Tu, J.; Suo, S.; Zhang, C.; Wong, K.; Urtasun, R. Towards Scalable Coverage-Based Testing of Autonomous Vehicles. In Proceedings of the 7th Conference on Robot Learning, PMLR, Atlanta, GA, USA, 6–9 November 2023; pp. 2611–2623, ISSN 2640-3498.
55. Hungar, H. A Concept of Scenario Space Exploration with Criticality Coverage Guarantees, Extended Abstract. In *Leveraging Applications of Formal Methods, Verification and Validation: Applications, 9th International Symposium on Leveraging Applications of Formal Methods, ISoLA 2020, Rhodes, Greece, 20–30 October 2020, Proceedings, Part III*; Lecture Notes in Computer Science; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; Chapter 19, Volume 12478, pp. 293–306. [CrossRef]
56. Gangopadhyay, B.; Khastgir, S.; Dey, S.; Dasgupta, P.; Montana, G.; Jennings, P. Identification of Test Cases for Automated Driving Systems Using Bayesian Optimization. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1961–1967. [CrossRef]
57. Khastgir, S.; Brewerton, S.; Thomas, J.; Jennings, P. Systems Approach to Creating Test Scenarios for Automated Driving Systems. *Reliab. Eng. Syst. Saf.* **2021**, *215*, 107610. [CrossRef]
58. Li, S.; Yang, J.; He, H.; Zhang, Y.; Hu, J.; Feng, S. Few-Shot Scenario Testing for Autonomous Vehicles Based on Neighborhood Coverage and Similarity. *arXiv* **2024**, arXiv:2402.01795.
59. Li, S.; He, H.; Yang, J.; Hu, J.; Zhang, Y.; Feng, S. Few-Shot Testing of Autonomous Vehicles with Scenario Similarity Learning. *arXiv* **2024**, arXiv:2409.14369.

60. Ken Mori, T.; Liang, X.; Elster, L.; Peters, S. The Inadequacy of Discrete Scenarios in Assessing Deep Neural Networks. *IEEE Access* **2022**, *10*, 118236–118242. [CrossRef]
61. Kaiser, B.; Weber, H.; Hiller, J.; Engel, B. Towards the definition of metrics for the assessment of operational design domains. *Open Res. Eur.* **2023**, *3*, 146. [CrossRef] [PubMed]
62. Design of Experiments (DoE). *Quality Management in the Bosch Group—Technical Statistics*; Robert Bosch GmbH: Stuttgart, Germany, 2010. Available online: https://assets.bosch.com/media/global/bosch_group/purchasing_and_logistics/information_for_business_partners/downloads/quality_docs/general_regulations/bosch_publications/booklet-no11-design-of-experiments-doe_EN.pdf (accessed on 21 October 2024).
63. *Scenario-Based Verification and Validation of Self-Driving Vehicles: Relevant Safety Metrics*; White Paper; Siemens Digital Industries Software & IVEX NV: Plano, TX, USA, 2022.
64. Certified Tester Foundation Level Syllabus. 2011. Available online: https://astqb.org/assets/documents/CTFL-2018-Syllabus.pdf (accessed on 16 October 2024).
65. Avizienis, A.; Laprie, J.C.; Randell, B.; Landwehr, C. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Dependable Secur. Comput.* **2004**, *1*, 11–33. [CrossRef]
66. Kuznietsov, A.; Gyevnar, B.; Wang, C.; Peters, S.; Albrecht, S.V. Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review *arXiv* **2024**, arXiv:2402.10086. [CrossRef]
67. Reich, J.; Hillen, D.; Frey, J.; Laxman, N.; Ogata, T.; Paola, D.; Otsuka, S.; Watanabe, N. *Concept and Metamodel to Support Cross-Domain Safety Analysis for ODD Expansion of Autonomous Systems*; Springer Nature: Cham, Switzerland, 2023. [CrossRef]
68. Parsons, J.E.; Mello, A.S. *Lecture Notes on Advanced Corporate Financial Risk Management-Chapter 5: Measuring Risk-Introduction*; MIT: Cambridge, MA, USA, 2010.
69. Zhan, W.; Sun, L.; Wang, D.; Shi, H.; Clausse, A.; Naumann, M.; Kummerle, J.; Konigshof, H.; Stiller, C.; de La Fortelle, A.; et al. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv* **2019**, arXiv:1910.03088.
70. Poggenhans, F.; Pauls, J.H.; Janosovits, J.; Orf, S.; Naumann, M.; Kuhnt, F.; Mayr, M. Lanelet2: A High-Definition Map Framework for the Future of Automated Driving. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018.