vehicles

MDPI

Review

Scenario Metrics for the Safety Assurance Framework of Automated Vehicles: A Review of Its Application

Erwin de Gelder [1,*], Tajinder Singh [2], Fouad Hadj-Selem [3], Sergi Vidal Bazan [4] and Olaf Op den Camp [1]

[1] TNO—Integrated Vehicle Safety, Automotive Campus 30, 5708 JZ Helmond, The Netherlands
[2] Siemens—Digital Industries Software, Automotive Campus 15, 5708 JZ Helmond, The Netherlands
[3] Vedecom—MobiLAB, 23 bis Allée des Marronniers, 78000 Versailles, France
[4] Applus + IDIADA—Electronics, Santa Oliva, P.O. Box 20, 43710 Tarragona, Spain
* Correspondence: erwin.degelder@tno.nl

Abstract

Ensuring the safety of Automated Driving Systems (ADSs) requires structured and transparent validation processes. Scenario-based testing has emerged as a widely adopted approach, enabling the targeted assessment of system behavior under diverse and challenging conditions. To offer a structured approach for scenario-based safety assurance, the European SUNRISE project developed the Safety Assurance Framework (SAF), which comprises stages such as scenario creation, allocation, execution, evaluation, decision-making, and in-service monitoring and reporting. Central to the SAF are scenario metrics, which quantify aspects such as coverage, criticality, and complexity and support evidence for safety cases. This paper provides a comprehensive overview of scenario-based scenario metrics relevant to ADS safety assessments. We categorize six core metric types: completeness, coverage, criticality, diversity/dissimilarity, exposure, and complexity. We explain their roles across the difference SAF components. This paper also discusses interdependencies among metrics, implementation challenges, and gaps where further research is needed, particularly in metric validation, aggregation, and standardization. By clarifying the landscape of scenario metrics and their application within the SAF, this work aims to support both practitioners and researchers in advancing scalable, data-driven safety assurance for ADSs.

Keywords: automated driving system; safety; assurance; scenario; metrics

Academic Editors: Tai-Jin Song and Yao Cheng

Received: 31 July 2025
Revised: 29 August 2025
Accepted: 8 September 2025
Published: 13 September 2025

Citation: de Gelder, E.; Singh, T.; Hadj-Selem, F.; Vidal Bazan, S.; Op den Camp, O. Scenario Metrics for the Safety Assurance Framework of Automated Vehicles: A Review of Its Application. *Vehicles* **2025**, *7*, 100. https://doi.org/10.3390/vehicles7030100

1. Introduction

To advance the deployment of Automated Driving Systems (ADSs) on public roads, ensuring their safety and reliability is crucial for both industry and regulators. Large-scale road testing is impractical for ADSs because the amount of testing that would be required to obtain enough statistical evidence of its safe operation is in the order of billions of kilometers [1]. Scenario-based safety validation is a widely supported approach in the automotive domain [2,3]. The scenario-based approach enables targeted evaluation of system behavior under diverse and challenging conditions. However, given the complexity of ADSs and the associated operating conditions they operate in, a structured and transparent approach for safety assurance is essential [4].

In response to the need for a structured approach for safety assurance, the European Safety assUraNce fRamework for connected, automated mobIlity SystEms (SUNRISE) project (https://ccam-sunrise-project.eu/, accessed on 12 September 2025) developed

the Safety Assurance Framework (SAF), which is based on the New Assessment/Test Method for Automated Driving (NATM) framework proposed by the UNECE [5]. The SAF structures the ADS evaluation into a series of processes that ultimately leads to a systematic argument for the safety case (more details on the SAF will follow in Section 4). The SAF is designed to be flexible towards the application that is under consideration (also referred to as the Subject Under Test (SUT)), the toolchains that are used throughout the assurance process, the operating conditions of the SUT, and external requirements that must be met, making it a valuable foundation for structured safety assurance.

An important aspect toward effective implementation of the SAF is the use of scenario metrics: quantitative measures that help evaluate the adequacy, performance, and relevance of the scenarios used throughout the safety assessment process. Metrics can guide scenario selection, support pass/fail evaluation, quantify test coverage, and provide evidence for the safety case. Despite their importance, the literature lacks a consolidated view of which types of metrics are relevant at each SAF stage and how they interrelate.

The goal of this paper is to fill this gap by providing a structured overview of scenario-based metrics relevant to the SAF. We categorize and describe key types of metrics such as scenario completeness, scenario coverage, scenario criticality, scenario diversity, scenario exposure, and scenario complexity. Furthermore, we discuss their roles across different SAF components. In doing so, we aim to support both practitioners and researchers in understanding the current state of the art, identifying areas for further development, and applying metrics effectively within the SAF or, more generally, any ADS safety assurance process.

This work is structured as follows: First, related surveys are briefly discussed in Section 2. Section 3 provides an overview of scenario metrics. The SAF is described in more detail in Section 4. This section also provides an overview of which metrics can be used at the different SAF components. This paper ends with a discussion and conclusions in Sections 5 and 6, respectively.

2. Related Surveys

Recent works have reviewed different perspectives of scenario-based assessments, which provide overviews of the different aspects of scenario-based assessments. For example, next to the works already mentioned in the introduction ([2–5]), Finkeldei et al. [6] introduce "Scenario Factory 2.0", a toolchain built around CommonRoad [7], for scenario-based testing. In [8], the use of large language models to enhance scenario-based testing workflows is reviewed. Yan et al. [9] propose a scenario generation framework that generates diverse scenarios with varying risk levels. While these works are not an exhaustive list, they already showcase that scenario-based testing methods rely on proper metrics. This observation further motivates our focus on consolidating and structuring scenario metrics.

Several prior works have surveyed or proposed metrics relevant to the assessment of ADSs, each with a particular focus or application domain. These contributions offer valuable insights into specific categories of metrics such coverage, criticality, and complexity. However, they often address these concepts outside the context of scenario-based evaluation. Therefore, this work should be seen as complementary to existing surveys, with the goal of providing a scenario-focused perspective on metrics relevant to the scenario-based safety assessment for ADSs.

For instance, Emran [10] provides a structured overview of data completeness measures, offering a classification of techniques used to quantify completeness in datasets. Although not tailored to driving scenarios, the conceptual foundation is relevant for understanding how completeness might be formalized in the context of scenario-based assessments of ADSs. Regarding criticality, Westhofen et al. [11], Wang et al. [12] provide a comprehensive review of surrogate safety measures that aim to measure safety without

the presence of safety-related events like collisions. For scenario complexity, Liu et al. [13] provide a detailed overview of metrics aimed at quantifying environmental, behavioral, and perceptual complexity in driving scenarios. Finally, the literature on coverage-based testing (e.g., structural coverage, requirements coverage, and parameter space exploration) is well-established, particularly in software and system testing (e.g., see [14]). However, overviews on coverage metrics related to driving scenarios with respect to an ADS' Operational Design Domain (ODD) are limited.

Despite the value of these works, there remain important gaps in the literature. Most notably, few surveys take a scenario-centric view when reviewing different types of metrics, such as completeness, criticality, and coverage, within the context of assessing the performance of ADSs against their ODDs. Moreover, there is limited guidance on how such metrics can be mapped to different stages of scenario-based assessments, such as scenario generation, test case allocation, and In-Service Monitoring and Reporting (ISMR). This article aims to address these gaps by providing a structured overview of scenario-based metrics specifically tailored to ADS safety assessments (Section 3). Moreover, we will highlight the role of these metrics within the SUNRISE SAF (Section 4). Figure 1 visualizes the position of this work, while aforementioned references provide an overview on either scenario-centric perspective on ADS assessments or specific metrics; this work provides an overview of metrics within the context of scenario-based assessments of ADSs.

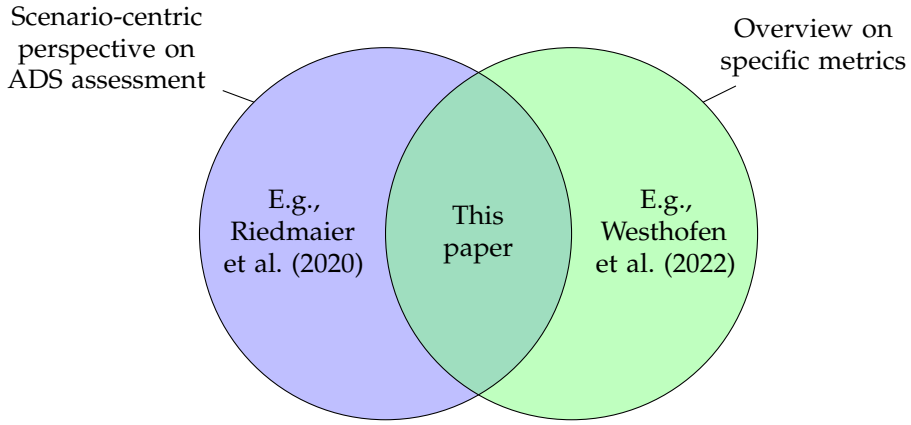

Figure 1. Positioning of this work: while there are several (review) works on scenario-centric perspectives of ADS assessments (e.g., Riedmaier et al. [2], but also [3–6,8,9]) and several works on specific metrics (e.g., Westhofen et al. [11], but also [10,12–14]), this work combines the two by providing an overview of metrics within the context of scenario-based assessments of ADSs.

3. Metrics

This section presents relevant metrics for scenarios that can be utilized when using scenarios within the SAF. While this section focuses on the metrics themselves, Section 4 will explain how the metrics can be used for the SAF. The metrics are grouped into six different types of metrics: scenario completeness (Section 3.1), scenario criticality (Section 3.2), scenario coverage (Section 3.3), scenario diversity and dissimilarity (Section 3.4), scenario exposure (Section 3.5), and scenario complexity (Section 3.6). These metrics are selected because they consistently emerge across the literature as key dimensions for evaluating scenarios and because each directly supports at least one stage of the SAF (see Section 4). For instance, the ISO 34502 standard [15] emphasizes risk- and function-oriented categories such as perception, planning, and control, which map closely to criticality and complexity. Similarly, the NATM framework proposed by the UNECE [5] stresses the importance of scenario coverage and representativeness, where the latter is closely related to scenario exposure. Though not explicitly mentioned, completeness is necessary for well-defined scenarios and thus relevant for both the NATM framework and ISO 34502 [15]. Diver-

sity and dissimilarity are mentioned in scenario clustering/selection studies, though no consolidated overview exists in the literature to the best of our knowledge. Together, these size metrics capture complementary aspects: completeness and coverage ensure adequacy of scenario descriptions and breadth of operational conditions; criticality and complexity address the level of challenge posed to the system; diversity/dissimilarity ensures non-redundancy within scenario sets; and exposure grounds the assessment in real-world likelihoods.

For each of the six different types of metrics, relevant metrics from the literature are briefly discussed. Table 1 provides an overview of these six types of metrics. In addition to these metrics, Section 3.7 presents other relevant types of metrics such as realism, rarity, and representativeness. The reason that these other types of metrics are not included in the list above is either due to their limited treatment in the existing literature or because they closely overlap with the six aforementioned metric types. These additional metrics are briefly introduced and discussed in relation to the six main metric types. Note that the following subsections can be read independently.

Table 1. Overview of different types of metrics for scenarios. In case of multiple definitions, they are separated by a semicolon.

Name	Definition in the Context of ADS Safety Assessment	Relevant References
Completeness	The extent to which a scenario description contains all the information necessary for meaningful analysis and decision-making	[10,16,17]
Criticality	Quantification of the potential risks and challenges in a scenario	[11,15,18–32]
Coverage	The adequacy of a testing effort; the extent to which a set of scenarios addresses a given ODD	[2,17,33–44]
Diversity or dissimilarity	Quantification of how two scenarios are different from each other; spread across a scenario set	[45–56]
Exposure	The likelihood of encountering a scenario	[57–67]
Complexity	The degree of challenge a scenario presents to a human driver or an ADS	[13,68–78]

3.1. Scenario Completeness

From all six core metrics, completeness metrics are the least discussed metrics in the literature (see also Figure 2). Many agree on the importance of completeness, but a commonly-agreed definition of completeness does not exist [10]. In [10], a list of various definitions of completeness is presented. In general, completeness refers to the state or degree of having all the necessary or appropriate parts. In the context of scenario-based safety assessments of ADSs, completeness indicates the extent to which a scenario description contains all the information necessary for meaningful analysis and decision-making. A complete scenario is one that is free from missing or ambiguous data, as well as specifies all relevant aspects—such as actor behaviors, environmental conditions, and time-based relations—required to simulate or evaluate the scenario accurately.

While completeness is often discussed alongside coverage, the two concepts serve distinct roles. Coverage, which will be more extensively discussed in Section 3.3 pertains to the breadth of scenarios—how well they span the ODD and the range of conditions under which the system is expected to operate. Completeness, on the other hand, is about depth—ensuring that each individual scenario is sufficiently specified. For example, a scenario might be incomplete if it lacks details such as vehicle velocities, road geometries, or environmental factors like lighting and weather. High coverage ensures that all relevant

types of situations are represented, but without completeness, those scenarios may be unusable for simulation, testing, or validation. Thus, completeness is essential for enabling confident assertions about system behavior under specific conditions.

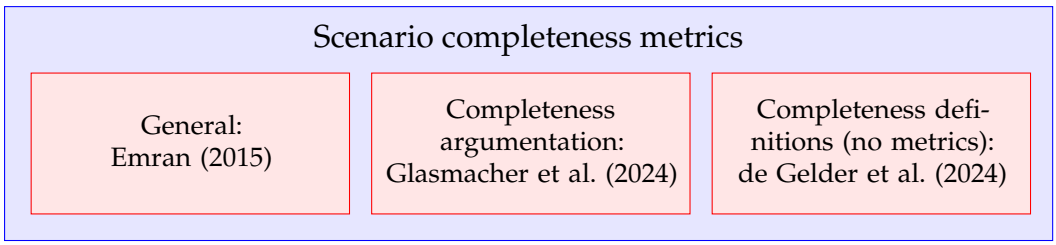

Figure 2. Overview of relevant studies for scenario completeness metrics, which focus on the general concept of completeness ([10]), completeness argumentation [16], or completeness definitions [17].

According to [16], several approaches mentioned in the literature aim to achieve higher completeness in the context of safety assessments of ADSs, but Glasmacher et al. [16] are the first to provide an argument for completeness. More specifically, they provide an argument for the state of completeness—which they regard as binary—of the so-called scenario concept, where scenario concept refers to a set of scenario categories, their definitions, and the relations between scenario categories. They define completeness for a use case "if all relevant driving situations are adequately captured", which is considered to be a binary state; i.e., completeness is either reached or it is not. It should be noted that completeness thus depends on the use case. Also, it is not further elaborated what "adequately" means.

In [17], the authors propose two different types of completeness that focus on concrete scenarios rather than the scenario concept. Since their use case is the development of a scenario database based on real-world data, the two different types of completeness focus on different aspects of this process. The first type of completeness addresses the question of whether "the driving data contain all relevant details of an ODD". The second type addresses the question of whether "the collected scenarios describe all relevant details that are in the driving data". Due to the word "relevant", the degree of completeness depends on what is considered to be relevant, which depends on the actual use case. For example, if a system's response depends on the color of the vehicle in front, the vehicle's color is considered relevant and must be described in order to reach completeness. Conversely, if another system's response does not depend on the vehicle's color, there is no need to describe this to reach completeness. In [17], providing a quantitative measure for completeness is left as future work.

One aspect of the completeness of a scenario description is the adherence to a specified format. For example, in the context of scenario description for virtual simulations, one of the most commonly utilized format is ASAM OpenSCENARIO XML [79]. Because a file specification exists, it is possible to ascertain whether a given scenario description adheres to the prescribed specification. Note that this does not imply that the content of the scenario description is sensible or complete. For example, it is straightforward to describe a scenario with two vehicles driving through each other or to simply leave out some important details while still adhering to the specified format. To check for such issues, a visual inspection can be conducted with a visualization tool of the scenario description. For ASAM OpenSCENARIO XML, Esmini (https://esmini.github.io/, accessed on 12 September 2025) could be used for that purpose.

In summary, the completeness of scenarios entails the degree of which all necessary information is contained in the scenario. This definition can be applied to concrete scenarios [17] as well as to a more abstract scenario concept [16]. As a start towards com-

pleteness measures, tools have been developed to check whether a scenario description contains the necessary information to be executable in a simulation environment.

3.2. Scenario Criticality

Criticality metrics are fundamental in evaluating and ensuring the safety and reliability of ADSs. These metrics quantify the potential risks and challenges in various traffic scenarios, offering a framework for assessing and mitigating hazards. It is important to note that while criticality metrics quantify the level of imminent danger or challenge within an individual scenario, they do not establish the overall safety of an ADS by themselves. In the context of human driving research, criticality metrics are often used as surrogate safety indicators to evaluate crash risk. Within ADS assessments, however, their role is to help identify and prioritize safety-relevant scenarios for testing. The broader evaluation of system-level safety, which integrates results across many scenarios and metrics, is addressed later within the SAF (Section 4).

As previously discussed in Section 2, a comprehensive list of criticality metrics is provided in [11]. Therefore, this section only briefly outlines the five categories of criticality metrics proposed by Cai et al. [18], and we refer the reader to [11] for a more thorough overview. For an overview of the literature reviewed below, see Figure 3. The five categories are as follows:

- Trajectory-based metrics: These metrics calculate the spatial or temporal gaps between traffic participants based on their trajectories or positions within a scene. Examples include Time Headway (THW) [19], gap time, distance headway, Time-to-Collision (TTC) [20], worst TTC [21], time to closest encounter [22], time exposed TTC [23], time integrated TTC [23], time to zebra [24], and post-encroachment time [25]. These metrics are crucial for scenarios where the precise movement and interaction of vehicles are central to assessing risk.
- Maneuver-based metrics: These metrics measure the difficulty of avoiding an accident through specific maneuvers such as braking and steering. For braking, key metrics include time to brake, deceleration to safety time, brake threat number [26], required longitudinal acceleration, and longitudinal jerk. For steering, important metrics include time to steer, steer threat number [26], required lateral acceleration, required longitudinal acceleration, and lateral jerk. These metrics are essential for evaluating the immediate actions required to prevent collisions.
- Energy-based metrics: These metrics assess the severity of a crash. For example, Yue et al. [27] use the kinematic energy of the ego vehicle to compute the scenario risk index. These metrics are critical for understanding the potential impact and damage severity in crash scenarios.
- Uncertainty-based metrics: These metrics capture the uncertainties inherent in traffic scenarios. The level of uncertainty in a scenario generally correlates with the number of challenges faced by the SUT. Examples include the pedestrian risk index by Cafiso et al. [28], which quantifies the temporal variation in estimated collision speed between a vehicle and a pedestrian, and the crash potential index [29], which estimates the average crash possibility if the required deceleration exceeds the maximal available deceleration. Schreier et al. [30] utilized Monte Carlo simulations to estimate behavioral uncertainties of traffic participants with the time-to-critical-collision probability. These metrics are pivotal for scenarios with high variability and unpredictability.
- Combination-based metrics: These metrics integrate several criticality metrics, addressing different aspects of a scenario to provide a more comprehensive assessment. Huber et al. [31] presented a multidimensional criticality analysis combining various metrics to evaluate overall scenario criticality. Baumann et al. [32] proposed a

combination-based metric that includes longitudinal acceleration, THW, and TTC. These metrics offer a holistic view but require careful consideration of the weights assigned to different components.

The diverse approaches to criticality metrics underscore the complexity and multi-faceted nature of traffic scenarios. Each class of metrics addresses specific aspects of risk, yet no single metric can be universally applied to all scenarios. Appropriate criticality metrics need to be tailored to the specific conditions of different scenarios, as a general and objective criticality metric for all scenarios does not yet exist.

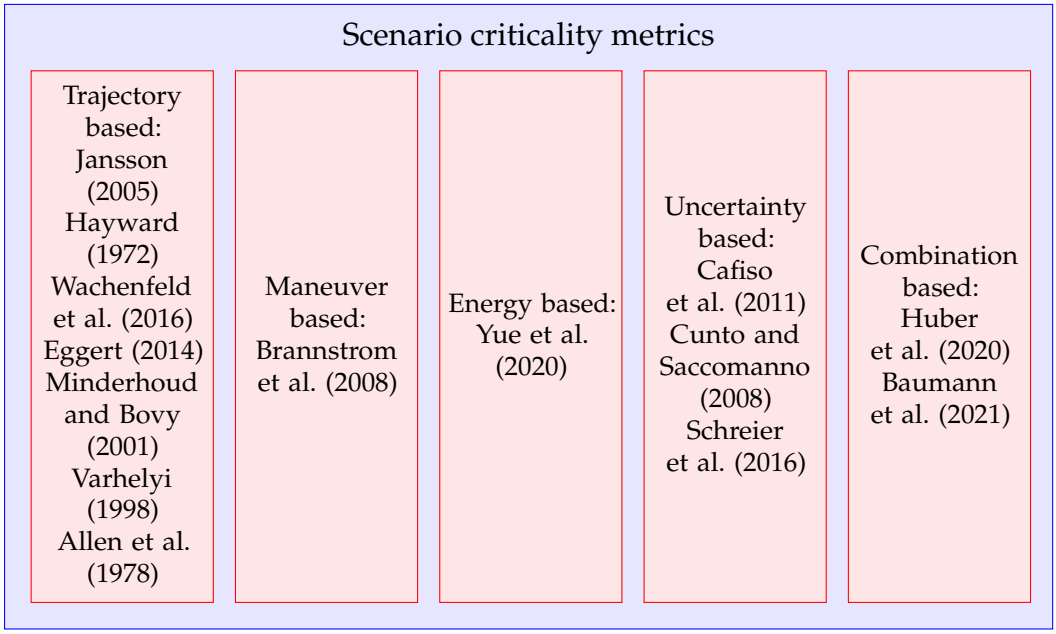

Figure 3. Overview of relevant studies for scenario criticality metrics, which focus on trajectories [19–25], maneuvers [26], energy [27], uncertainty [28–30], or combinations thereof [31,32].

3.3. Scenario Coverage

Coverage generally refers to the degree to which something deals with something else. In the field of software engineering, coverage is a measure of the verification progress [33]. Since there are multiple ways to measure (the degree of) verification completeness, Piziali [33] argues that there is no single (best) way to define coverage. For example, in the application of software engineering, coverage can be related to the fraction of functional requirements that have been addressed (functional coverage), the fraction of the code that has been executed during the verification process (code coverage), and the fraction of assertions that have been evaluated (assertion coverage). For the different types of coverage, multiple measures can be considered. For instance, code coverage can be measured in terms of the lines of codes that have been executed, the branches that are covered, etc.

In [34], the authors highlight the importance of coverage metrics in testing autonomous vehicles. They argue that inadequate coverage of potential situations an autonomous vehicle might encounter is similar to insufficient testing. To address this, Alexander et al. [34] propose a "situation coverage metric". They suggest that this metric should be tractable with the following characteristics:

- Calculable percentage: The metric should be expressible as a percentage. Metrics like the number of kilometers driven or the number of (simulated) scenarios are inadequate because they can be infinite. Similarly, the number of failures found is not useful since the total number of possible failures is unknown.

- Coverage of 100% achievable: The metric should allow for 100% coverage to be realistically achievable under practical conditions.

In the context of testing ADSs, "coverage" is frequently used to assess the adequacy of a testing effort and to determine when testing can be concluded [35]. Riedmaier et al. [2] described "scenario coverage" as the extent to which the concrete scenarios used for testing encompass the entire scenario space, though they did not provide specific quantitative measures. Traditional metrics, such as requirement and code coverage [33], are also relevant for ADSs. Additionally, specific coverage metrics have been developed for Automated Driving (AD). This section will highlight several of these metrics, which can be categorized based on the two aforementioned properties identified by Alexander et al. [34]: metrics that cannot be expressed as a percentage, metrics that can be expressed as a percentage but for which 100 % is not realistic, and metrics that can be expressed as a percentage for which 100 % is achievable. An overview is shown in Figure 4.

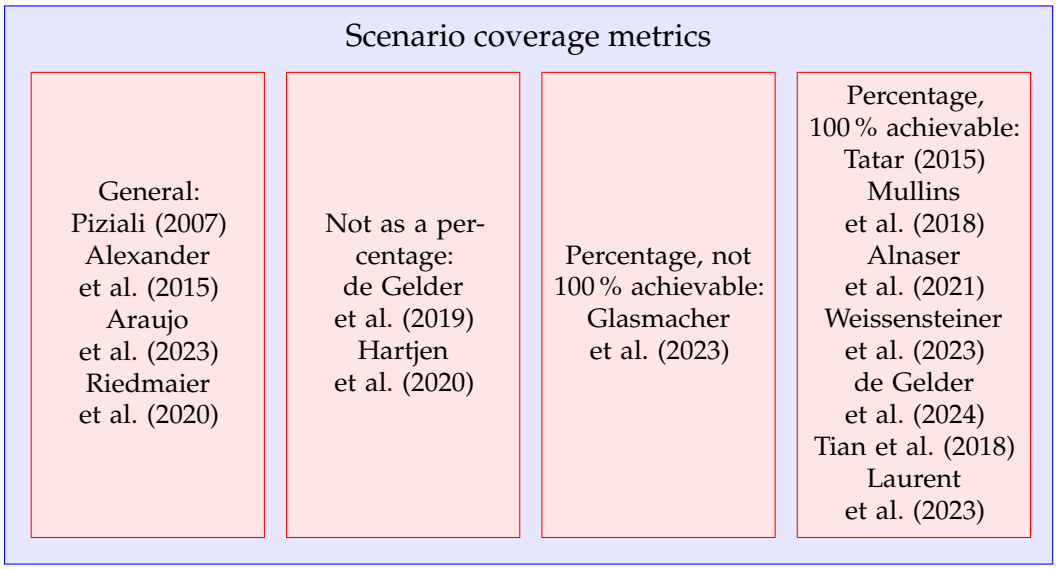

Figure 4. Overview of relevant studies for scenario coverage metrics, which focus either on the general concept of completeness [2,33–35] or can be categorized as metrics not as a percentage [36,37], metrics as a percentage but without 100% achievable [38], and metrics as a percentage with 100% achievable [17,39–44].

In [36], a measure for the uncertainty of an estimated Probability Density Function (PDF) is used to measure the degree of completeness of the acquired data. The used measure ranges from zero to infinite, so it cannot be expressed as a percentage. Zero is only reached in case infinite data is used to estimated the PDF. In [37], two measures are proposed for measuring the "saturation effect" in recorded data. For the maneuver layer, the Kullback–Leibler divergence between the PDF estimated with all data and the PDF estimated with less data is used. This provides a number ranging from zero to infinite, with zero if and only if the two PDFs are equal. For the behavior layer, the number of unique maneuver sequences is used. The idea is that this number should reach an asymptote. However, because this asymptote is unknown, this number cannot be expressed as a percentage. Because the metrics in [36,37] cannot be expressed as percentages, it is difficult to determine a threshold for which the data can be regarded as sufficient or saturated.

Glasmacher et al. [38] proposed a coverage metric based on the values of the scenario parameters. Here, the parameter could refer to scenario-level parameters, like parameters related to the road geometry, environmental conditions, etc., as well as behavioral parameters of traffic participants, such as initial speed, deceleration, etc. In their approach, each scenario is represented by an ellipsoid in a parameter space. The total covered space is

represented by the union of all ellipsoids. The degree of coverage is calculated by dividing the total covered space by the space that can potentially be covered, where the latter is estimated based on the assumption that the covered space as a function of the total number of scenarios can be represented by a cumulative Weibull distribution function. As a result, the coverage can be expressed as a percentage. However, since a cumulative Weibull distribution function is assumed, 100% coverage is only achieved if an infinite amount of data is used.

A number of coverage metrics from the third category, i.e., metrics that can be expressed as a percentage for which 100% is realistically achievable, are presented in the literature. In [39], the so-called state coverage is proposed, which is a percentage of predefined states that have been reached during testing. Similarly, in [40], the coverage is expressed as the percentage of regions covered by the robot. Both these methods require us to determine the states or regions that must be covered during testing before the actual testing takes place.

A framework for the coverage of scenes, i.e., a description of the environment at a certain point in time, is presented in [41]. Here, a method is proposed to discretize scenes such that they can be enumerated. Once enumerated, the percentage of scenes covered during testing can be calculated. However, the details to reproduce the metric of [41] are missing, and so far, no practical results limiting the use of the presented method are presented.

In [42], the coverage of an ODD is calculated by breaking the ODD down into predefined logical scenarios. It is assumed that the predefined logical scenarios fully cover the ODD. The coverage of an individual logical scenario is based on the coverage of the concrete scenarios that are covered by the logical scenario. Similarly, the coverage of a concrete scenario is based on the coverage of the so-called continuous parameters that are part of the concrete scenarios. Note that the continuous parameters of a scenario are considered individually, meaning that the coverage of [42] does not consider different combinations of parameter values.

Also in [17], metrics for the scenario-based coverage are proposed. Two type of coverage metrics are distinguished. The first type considers the coverage of all relevant aspects of an ODD. By encoding these relevant aspects using tags, e.g., following the ISO 34504 standard [80], the coverage is determined by the number of scenarios containing the predefined tags. The second type of coverage metrics considers the extent to which the collected scenarios cover all relevant aspects that are in a data set. One metric calculates the percentage of time instants that are covered by n scenarios. A second metric computes the percentage of relevant actors that are covered by the scenarios, where an actor is deemed relevant based on some predefined rules. A third metric combines the other two and calculates whether all actors are covered at the time instances at which these actors are considered relevant.

Other coverage metrics related to the (testing of) AD focus on the internal state of an ADS. For example, in [43], the authors focused on the Deep Neural Network (DNN) of an ADS. They proposed the so-called neuron coverage, which is the ratio of activated neurons during all tests to the total number of neurons of the neural network(s). In [44], it is assumed that an ADS computes its decisions using parameterized rule-based systems and cost functions, meaning that parameters characterize the decision process. They proposed "parameter coverage", where a scenario covers a parameter if changing the parameter's value with a certain amount leads to different simulation results.

3.4. Scenario Diversity and Dissimilarity

The scenario dissimilarity metric compares two scenarios to identify how different they are from each other. The scenario diversity metric extends the notion of dissimilarity to a set of scenarios to measure the spread across the set. These metrics may be applied at any scenario abstraction stage, whether functional, logical, or concrete. The metrics have various applications for scenario-based testing, including

- Identifying redundant scenarios such that they can be skipped to reduce test efforts.
- Clustering and categorization of concrete scenarios to obtain logical scenarios or scenario categories. Logical scenarios/scenario categories help with understanding, storage, and querying of scenarios.
- Promoting a diverse set of scenarios when using scenario generation methods such as optimization.

Existing dissimilarity metrics for scenarios can be broadly classified into three categories (for an overview, see Figure 5):

1. Dissimilarity based on scenario parameters: These metrics are applied particularly to multiple concrete scenarios of the same logical scenario. As concrete scenarios are obtained by sampling values for parameters of the logical scenario, dissimilarity is computed by comparing parameter values of concrete scenarios. For example, Zhu et al. [45] compute dissimilarity based on the Euclidean distance in parameter space. Alternatively, Zhong et al. [46] define the dissimilarity of a (traffic violation) scenario based on the percentage of scenario parameters that differ between two scenarios. Here, a continuous parameter is said to differ between two scenarios when the difference in the parameter value is greater than a user-defined resolution.
2. Dissimilarity based on scenario trajectories: These metrics compute dissimilarity considering the complete trajectories of all actors in each scenario. For example, Ries et al. [47] use dynamic time warping to estimate similarity between the trajectories of actors in two scenarios. Nguyen et al. [51] use the Levenshtein distance to compute the similarity between trajectories. The Levenshtein distance measures the number of "edits" needed to convert one trajectory to another. Alternatively, Lin et al. [48] create matrix profiles that consist of dissimilarities between the subsequences of one trajectory with the nearest neighbor sub-sequences from the other trajectory. The dissimilarity is based on the number of elements that are lower than a certain threshold.
3. Dissimilarity based on scenario features: These metrics define dissimilarity based on features extracted based on expert knowledge or through feature extraction methods. The considered features include, e.g., behavior of scenario actors (e.g., average occupancy around the ego vehicle) and ODD features (e.g., road layout orientation). Kerber et al. [49] compute average occupancy of an 8-cell grid around the ego vehicle over the entire scenario and use it as a dissimilarity measure to compare scenarios. Kruber et al. [50] perform unsupervised random forest clustering based on road infrastructure and trajectory features and use hierarchical clustering to estimate a similarity measure. Alternatively, in [51,52], feature maps are computed based on similar features, including behavior aspects such as steering angle standard deviation. Some studies prioritize the critical segments of scenarios for dissimilarity calculation. Wheeler and Kochenderfer [53] determine the critical segment based on a risk threshold and then estimate dissimilarity based on behavioral features such as relative speeds, acceleration change, and attentiveness. References [54,55] use criticality metrics, e.g., the scene of the minimum distance between actors, to determine the most critical scene. The dissimilarity score is based on both discrete features, such as

driving path ids and actor types, and continuous features—which characterize the interaction, e.g., the relative heading angle of actors. A case study is presented on a database with a thousand scenarios generated by simulations in an intersection.

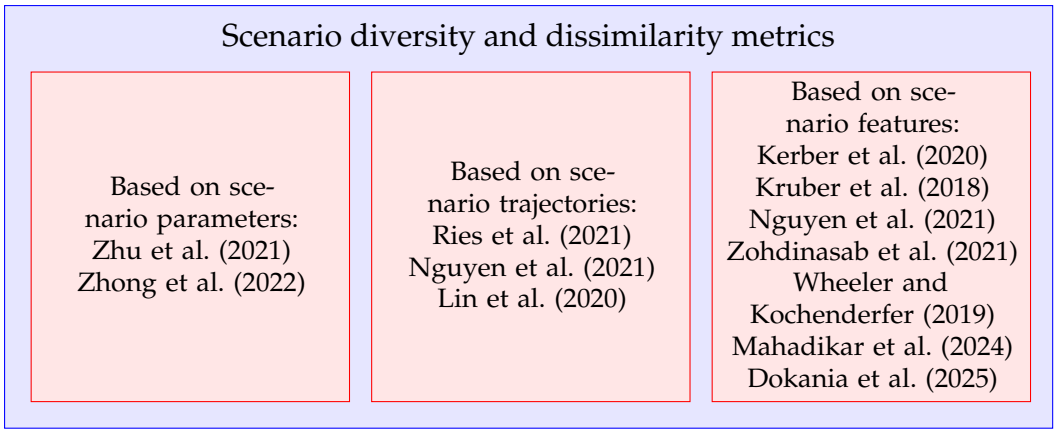

Figure 5. Overview of relevant studies for scenario diversity and dissimilarity metrics, which focus on scenario parameters [45,46], scenario trajectories [47,48,51], or scenario features [49–55].

Computing dissimilarity based on scenario parameters is straightforward and efficient, with low computational load and no additional processing of scenarios. However, this dissimilarity measure is independent of whether the two scenarios would pose different challenges to the ego vehicle (system under test). It is not distinguished which scenario parameters influence a safety-critical interaction with the ego vehicle, and how this influence changes in different regions of the parameter space. In contrast, dissimilarity based on trajectories captures the changes in interactions of other actors with the ego vehicle. However, complete trajectories are needed for computing dissimilarity, leading to increased computational load. Furthermore, unnecessary information may skew the dissimilarity metric, for example, when the trajectories of actors are much earlier than the actual, safety relevant interaction with the ego vehicle.

The third group of methods, dissimilarity on features, benefits from the ability to emphasize relevant features of the scenario as defined by experts or by data-driven extraction. Features extend beyond trajectories to consider the ODD and other behavioral aspects. In addition, specific segments of scenarios may be considered, for example, when part of the scenario after a criticality threshold for the ego vehicle is exceeded. Thus, the dissimilarity measure includes the notion of safety relevance and can be fine-tuned based on a given use case. Depending on the chosen features, it may be necessary to perform testing to obtain the features, for example acceleration change during the critical segment as in [53].

The diversity of a scenario set is established by extending the dissimilarity measure to the entire data set, providing a measure to quantify average dissimilarity and spread. The research on diversity metrics for scenario-based testing is still limited. Tian et al. [56] measure the average dissimilarity of a new scenario from an existing set. The average dissimilarity is used as an indicator of the increase in diversity due to the scenario. Alternatively, Zohdinasab et al. [52] map scenarios to certain cells within a feature map based on feature values. Then, they measure diversity using a sparseness measure, which is defined as the average maximum Manhattan distance between the occupied cells in the feature map.

3.5. Scenario Exposure

Scenario exposure metrics are related to the frequency, time spent, or distance traveled in a specific driving scenario in the real world. These metrics encompass aspects such as the

scenario probability of specific scenarios, the uncertainty associated with this probability, and the ability to predict and prepare for future scenarios. In practice, scenario exposure is typically expressed using the scenario probability, which itself is commonly estimated. It can be useful to also consider the uncertainty of the estimation. Hence, scenario probability uncertainty is also considered.

As part of the exposure, this section also contains a description of existing metrics for scenario foreseeability. This notion is mentioned in the United Nations Regulation 157 (UN R157), which requires that the ADS avoids any collisions that are reasonably foreseeable and preventable. Thus, it is required to determine all scenarios that are reasonably foreseeable. The following subsections will treat the different aspects of scenario exposure for which an overview is presented in Figure 6.

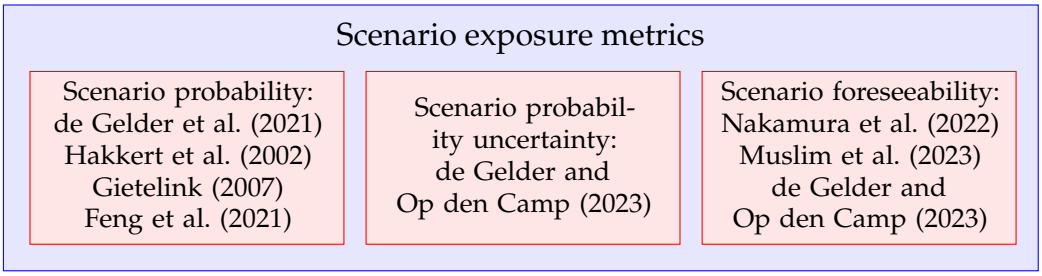

Figure 6. Overview of relevant studies for scenario exposure metrics, which focus on the scenario probability [57–59,61], scenario probability uncertainty [60], or scenario foreseeability [65–67].

3.5.1. Scenario Probability

Several methodologies have been developed to estimate scenario exposure using Naturalistic Driving Data (NDD) and Field Operational Test (FOT) data, each with its own advantages and limitations. For scenario exposure, a distinction is typically made using a "type of scenario", such as an "abstract scenario" [81] and a "scenario category" [82], and the parameter values within such a "type of scenario". For the former, an expectation of the number of encounters of such a situation can be determined, e.g., the expected number of individual scenarios belonging to a certain type of scenario within one hour of driving or within a certain predefined distance. For the latter, such an expectation is typically meaningless because the probability of encountering a certain scenario with those particular parameter values is zero. In that case, it is more useful to consider the probability density of the scenario parameters.

In the ISO 26262 standard [83], the exposure of being in a certain operational situation is qualitatively defined. The highest exposure (E4) is used if the situation is almost certain to happen during a single drive. E3, E2, and E1 are used for medium probability, low probability, and very low probability, respectively, where each class differs in one order of magnitude. The exposure classification E0 is used to indicate that a situation is considered incredible.

Expressing the exposure qualitatively supports further analysis of risks, but a—possibly more precise—quantitative expression of the exposure provides more possibilities for further analysis. Regarding the exposure of different types of scenarios, de Gelder et al. [57] have expressed the exposure as the expected number of encounters per unit of time for scenarios within a specific scenario category. Their work relies on real-world driving data, such as the data set from Paardekooper et al. [84], which includes 6000 km of public-road driving. This data-driven approach provides a robust basis for estimating exposure frequencies and identifying critical scenarios.

Hakkert et al. [58] have defined exposure within the context of road safety, focusing on various measures such as the number of kilometers traveled, time spent in traffic, and traffic volumes at intersections. These measures offer a practical way to quantify exposure

but often require extensive and high-quality data, which can be challenging and expensive to collect.

Regarding the exposure at the parameter value level, this comes down to either assuming a particular PDF or estimating a PDF based on some observations. Methods to estimate a PDF can be divided into two groups: parametric density estimation and non-parametric density estimation. With parametric density estimation, a particular shape of the PDF is assumed, while the corresponding parameters are estimated based on the data, e.g., by maximizing the likelihood of the samples. In the domain of scenario-based assessments for AD, Gietelink [59] assumed a Gaussian distribution of the scenario parameters. With the increase in data, more sophisticated (but data-hungry) methods could be employed when estimating the probability densities, such as kernel density estimation [57,60]. These methods, however, generally scale badly with an increasing number of parameters, which is why it is not uncommon to assume that the parameters are independent (see, e.g., [61]).

3.5.2. Scenario Probability Uncertainty

Despite the importance of the uncertainty of estimated probabilities, this has not been discussed often in the literature in relation to scenario exposure. However, outside the field of automated driving, extensive studies on this topic are available. Here, two different approaches can be distinguished:

- With the first approach, a parametric distribution is used to estimate the PDF, such as a normal or Gaussian distribution or a gamma distribution. In those cases, the distribution parameters (not to be confused with the scenario parameters for which the PDF is estimated) are typically fitted to some data. When using a Bayesian approach to fit those distribution parameters, the posterior uncertainty of the distribution parameters can be used to estimate the uncertainty of the density [62].
- With the second approach, a non-parametric distribution is used to estimate the PDF, such as Kernel Density Estimation (KDE). In those cases, the uncertainty is either based on a theoretical model or bootstrapping is used [63]. In the domain of AD, bootstrapping is used in [36,60] to estimate the probability uncertainty of the scenario parameters' probability density.

3.5.3. Scenario Foreseeability

Regulations for the type approval of ADSs require that the activated system does not cause any collisions that are reasonably foreseeable [64]. To determine what scenarios are considered to be "reasonably foreseeable", one can look at the PDF of the parameters and consider the parameter values at the "edges" to be not reasonably foreseeable. Nakamura et al. [65] exploited this idea to determine the "reasonably foreseeable" range of parameter values. Their approach assumes that scenario parameters are independently distributed according to the Beta distribution. From this, a parameter range capturing 99 % of the distribution is calculated, and all these parameter values are considered to be reasonably foreseeable. This analysis is applied in [65] to cut-in scenarios. In an extension, a similar analysis is performed for cut-out scenarios in [66].

This approach is expanded in [67], where two alternative methods are proposed to estimate "reasonably foreseeable" parameter values. Their first method employs non-parametric KDE, allowing the PDF to adapt to the data without assuming parameter independence. Their second approach utilizes extreme value theory, applying the generalized Pareto distribution to model extreme parameter values. These methods are demonstrated through case studies involving scenarios from [65] and an additional scenario where the ego vehicle approaches a slower vehicle.

3.6. Scenario Complexity

In the literature, there is no universally accepted definition or metric, and multiple approaches have emerged depending on the application context. In [68], the complexity of a system is related to the difficulty to predict the behavior, while Issler et al. [69] define scenario complexity as the randomness of the scenario. More generally, scenario complexity is typically understood as the degree of challenge a scenario presents to a human driver or an ADS and is often influenced by more than one factor, such as the number of elements and dynamic actors, the variety of the elements and actors, the behavior of the actors, and the relation between the elements and actors [70]. Figure 7 provides an overview of different approaches to measure scenario complexity.

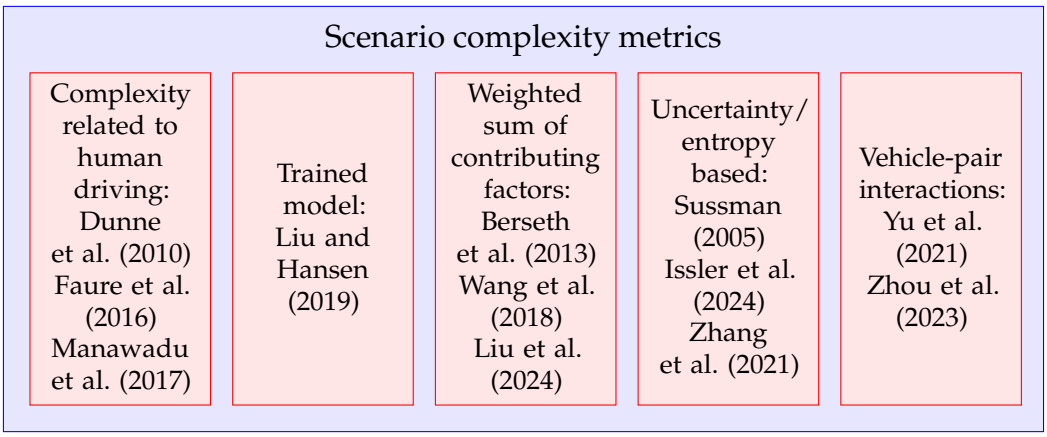

Figure 7. Overview of relevant studies for scenario complexity metrics, which are based on complexity related to human driving [71–73], a trained model [74], a weighted sum of contributing factors [13,75,76], uncertainty/entropy [68,69,77], or vehicle-pair interactions [70,78].

Research focusing on the challenge a scenario presents to a human driver typically expresses scenario complexity using the estimated complexity of the task of the human driver to deal with a certain traffic situation. For example, in [71], scenario complexity is based on the following three factors:

- The complexity of the task, i.e., the number of acts that the driver needs to perform;
- The number of possible ways the task can be performed, meaning that the driver need to take more decisions if there are more ways to perform a task;
- The number of external stimuli.

In [71], these three factors are combined by taking the sum of the first two factors and multiplying this with the third factor, leading to a single number that quantifies the scenario complexity. Other examples of measures for scenario complexity related to human drivers are presented in [72], which bases the complexity on the richness of the driving environment, and [73], which bases traffic complexity on the traffic density.

Given that the complexity of a scenario is often based on several factors, there are two main approaches used in the literature to combine the different factors. The first approach is using a trained model. For example, in [74], scenarios are labeled based on the perceived complexity, and a model is trained using a random forest so that the model can predict the perceived complexity of a scenario. For the features that are used by the model to predict scenario complexity, Liu and Hansen [74] use environmental information extracted from OpenStreetMap, surrounding vehicle information derived from video, and prior environmental knowledge such as weather, time, and driving location. The second approach is determined by taking the weighted sum of the different factors [13,75,76]. In [13], scenario complexity is determined by combining factors related to the environment (weather, illumination, daytime, or night-time), road (obstacles and road condition), and

dynamic entities (type and occlusion level). This approach offers a highly flexible metric, although determining the appropriate weights may not be straightforward.

To express the challenge of a scenario for an ADS, several works address the complexity of the dynamic part. An increasingly prominent approach is based on information theory and machine learning. For example, in [77], information entropy is used to express the uncertainties of all dynamic entities, which are used to express scenario complexity. Based on this, in [69], a framework that leverages entropy-based metrics to quantify the unpredictability and variability of the surrounding agent behavior is proposed, directly linking scenario complexity to the decision-making challenge for an ADS. In [70,78], a three-step approach is used to determine the complexity of the dynamic part. First, vehicles that are part of the so-called dynamic influencing area of the ADS are selected. Second, vehicle-pair complexity is computed based on the encounter angle, relative velocity, and relative distance. Third, a single quantity is obtained by integrating the vehicle-pair complexities over all pairs and after applying some form of smoothing. This approach has been shown to be consistent with complexity ratings of human drivers [70].

3.7. Other Metrics

This section briefly discusses other types of metrics. As mentioned before, the following metrics are not extensively discussed either due to their limited treatment in the existing literature or because they closely overlap with any of the six metric types that are discussed above. The relation with the metrics discussed above will be highlighted.

3.7.1. Realism

Since the use of virtual simulations is inevitable for the assessment of (high-level) ADSs, the development of high-fidelity simulations has received considerable attention. Many efforts have been put into reducing the so-called sim2real gap, as the extent of the sim2real gap can have a large influence on the evaluations of ADSs [85]. The quantification of the real2sim gap typically focuses on the gap between the model of the SUT [86,87], the gap between data generated by sensors [88,89], and the gap between the resulting behavior of an SUT [90,91]. These aspects of the real2sim gap go beyond the scenario descriptions themselves, which is why this work does not provide a further review on metrics addressing these aspects. Another contributor to the sim2real gap is the limited description of a scenario compared to the details in the real world. For example, even though background in a camera image can be influenced by the color of the surrounding buildings, not all of these colors may be described as part of a scenario. Following [17], this contributor to the sim2real gap is addressed by scenario completeness metrics (Section 3.1).

3.7.2. Rarity/Novelty

The rarity of a scenario can be expressed using metrics related to scenario exposure (Section 3.5). Although there is no clear definition of "novelty", one might argue that a "novel scenario" should be both rare and distinct from already known scenarios. Therefore, in addition to scenario exposure metrics, scenario diversity and dissimilarity metrics (Section 3.4) could be utilized to express novelty.

3.7.3. Reproducibility

Reproducibility is a critical aspect of scenario-based testing for ADSs. Achieving reproducibility requires that scenarios are described in such a way that they minimize ambiguity and interpretation errors. Completeness metrics (Section 3.1) help to assess whether all necessary parameters, constraints, and contextual elements (e.g., actor behaviors, road geometry, and weather conditions) are explicitly defined, minimizing ambiguity and interpretation errors. Variations in scenario execution may also arise due to stochastic

elements. Dissimilarity metrics (Section 3.4) can be used to quantify differences between multiple executions of what is nominally the same scenario.

3.7.4. Outcome Severity

Outcome severity metrics quantify the consequences of a scenario, such as the impact speed in the event of a collision and the likelihood of a resulting injury or damage. These metrics often overlap with scenario criticality metrics (Section 3.2), but they measure different aspects of risk: Scenario criticality metrics focus on the urgency of a conflict (e.g., using TTC), whereas outcome severity metrics address the impact if such a conflict is not avoided. Outcome severity metrics, such as the maximum abbreviated injury scale [92], are more related to the vehicle under test, whereas severity metrics are typically related to a scenario, which is why this paper put more emphasis on the latter.

3.7.5. Traceability

Traceability refers to the ability to track the origin of each scenario and any changes to it. In the context of ADS assessments, traceability plays a critical role in ensuring transparency, consistency, and accountability. Traceability could help with quantifying other metrics. For example, it may allow for reasoning of the ODD in which a certain scenario is encountered, thereby helping to quantify the coverage (Section 3.3) of an ODD as well as estimating the exposure (Section 3.5) of the scenario within an ODD. Typically, aspects of traceability are not quantified, which is why this work does not further elaborate on this topic. Instead, traceability is typically supported by qualitative attributes, structured metadata, and auditability criteria.

3.7.6. Representativeness

Scenario representativeness refers to how well a given scenario or a set of scenarios reflects the conditions and situations the ADS is expected to encounter in its ODD. When referring to a set of scenarios, representativeness refers to the extent to which the (relevant) characteristics of the scenarios reflect the characteristics of scenarios within a specific ODD. In [93], a "scenario representativeness metric" that compares a set of generated scenarios with a set of observed scenarios is proposed. They accomplish this by measuring the discrepancy of the parameter distributions using the Wasserstein metric of the two different scenario sets.

When referring to a single scenario, a representative scenario is one that is realistic, relevant, and frequent enough to support meaningful conclusions about the system's safety and performance in its intended use. To the best of our knowledge, no existing study in the field of automated driving explicitly defined a metric under the term "representativeness" for single scenarios. A reason for this could be that it is closely linked with scenario exposure: scenarios that are highly unrealistic and/or infrequent are inherently associated with low scenario exposure (Section 3.5) values.

4. Metrics for the Safety Assurance Framework

The Cooperative, Connected, and Automated Mobility (CCAM) Safety Assurance Framework (SAF) is the main deliverable of the SUNRISE project [94]. The SAF is designed to accelerate the safe deployment of ADSs. (To be more precise, the SAF targets CCAM systems. In this work, we refer to ADSs as the core component of broader CCAM systems, which also include vehicle connectivity and cooperative functions.) It aims to create a demonstrable positive impact towards safety.

The SAF is schematically shown in Figure 8. This figure also indicates where the different types of metrics can contribute to the SAF. The SAF is based on the NATM document from United Nations' world forum for harmonization of vehicle regulations [5].

The input of the SAF are the ODD, requirements related to the system behavior, external requirements, and test objectives. Next to the input, there are five main components that can be distinguished:

1. *Scenario*: It creates, formats, and stores (test) scenarios in databases;
2. *Environment*: It converts scenarios into concrete test cases and runs them on various testing environments;
3. *Safety Argument*: It evaluates test results, coverage, and overall system safety, which leads to a decision on a pass or fail for the SUT;
4. *In-Service Monitoring and Reporting (ISMR)*: It monitors the system during deployment, ensuring continual safety and providing input for future system designs;
5. *Audit*: It ensures proper safety processes throughout the development lifecycle.

The following five subsections will provide more details on these five main components as well as how the metrics of Section 3 can contribute to the SAF components.

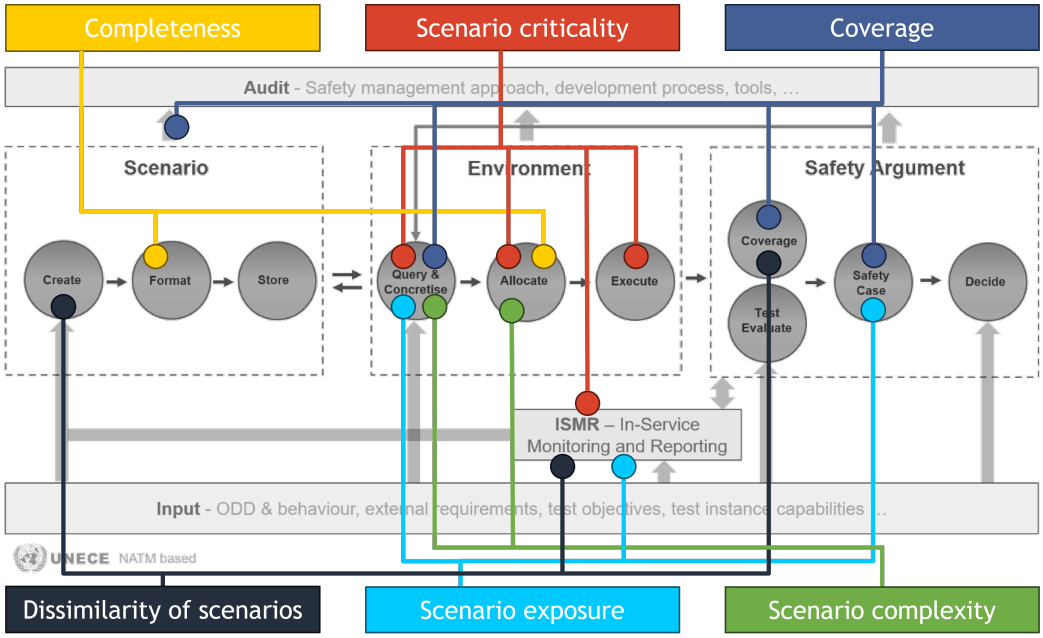

Figure 8. Safety Assurance Framework workflow including an overview of where the different types of metrics can contribute to the SAF.

4.1. SAF Component—Scenario

The *scenario* component consists of three parts: *create*, *format*, and *store*. Multiple approaches are possible and even desired for complementary reasons. Multiple SCenario DataBases (SCDBs) result from this process, and through the SUNRISE data framework that accesses these SCDBs, users of the SAF can obtain scenarios from various sources.

The first part of the *scenario* component (*create*) concerns the generation of scenarios, e.g., using data-driven and knowledge-driven approaches. A data-driven approach could be used to extract scenarios from real-world driving data such as methodologies like StreetWise [3]. Scenarios may also be created on the basis of system requirements, where the scenarios are intended to verify the conformance to the requirements. Note that at this stage, the scenario dissimilarity metrics might be useful for the creation of scenarios. These metrics could identify redundant or near-duplicate scenarios such that they can be skipped to reduce test effort and optimize databases.

The second part (*format*) involves the formatting of the scenarios such that these can be stored into a SCDB. Describing the scenarios into a computer-readable format facilitates easy access, interpretation, and integration of the scenario data across different systems

and tools. When dealing with test scenarios, a common format is ASAM OpenSCENARIO XML [79]. To describe scenarios observed in real-world driving data, a structured format for traffic recordings [95] or a format based on an object-oriented framework [82] might be used. At this stage, it can already be checked whether the formatted scenarios contain all relevant data, which is why completeness metrics can be applied at this stage.

After scenarios are created and formatted, the next step is to (*store*) them in a SCDB. The implementation of the storage is up to the SCDB owner, but requirements are set out regarding the interface of the SCDB with the subsequent components through the SUNRISE data framework. Although all metrics could be part of the metadata of scenarios and stored as such in the SCDBs, there is no direct use of these metrics for the storage itself.

4.2. SAF Component—Environment

The scenarios within the *store* component are the main input to the *environment* component. These scenarios are retrieved through the SUNRISE data framework and, together with the overall SAF input, which is utilized to create test cases that are executed in an allocated test environment. This component consists of three parts: *query and concretize*, *allocate*, and *execute*.

The *query and concretize* part is responsible for querying scenarios from the SCDB and defining concrete test scenarios. Together with the test objectives and pass/fail criteria, test cases are formulated. For the purpose of defining the test scenarios, many different approaches can be used [96]. For the different approach, different metrics could be utilized. One approach is to focus the test effort on the more critical or complex scenarios; thus the scenario criticality and scenario complexity metrics could be used. In addition, it is typically desired to use test scenarios that are representative of the ODD, and scenario exposure metrics could be of use to define whether scenarios are representative. Furthermore, the test scenarios should cover the ODD, which is why coverage metrics are also relevant for this part.

In the next step, *allocate*, test cases are assigned to appropriate testing environments, such as hardware-in-the-loop, proving grounds, or virtual simulations [97]. Each environment requires different levels of scenario detail and fidelity. Completeness metrics, which indicate whether a scenario contains all necessary and relevant information (for a specific testing environment), can guide this allocation process. Additionally, scenario criticality may influence the choice of environment: for example, a scenario with a high criticality score might be better suited for execution in a controlled or safe environment, such as a virtual simulation, to minimize risk during testing. Lastly, scenario complexity can be an important factor in choosing the testing environments. For instance, highly complex scenarios may not be feasible in all testing environments.

In the *execute* step, the selected scenarios are run in their designated testing environments, and the performance of the ADS is systematically evaluated. This involves monitoring key outputs, such as safety margins, rule compliance, and system responses. Scenario criticality metrics can play a valuable role in interpreting the test outcomes by providing context for how challenging or safety-relevant a scenario is. For example, an ADS's behavior in high-criticality scenarios may warrant closer scrutiny, as these situations often represent edge cases or conditions with a high potential for failure.

4.3. SAF Component—Safety Argument

The *safety argument* component evaluates an ADS's safety through four stages: *coverage*, *test evaluation*, the *safety case*, and *decide*. This component uses the test results that follow from the *environment* component together with the input, such as the ODD, behavioral requirements, and external requirements, to determine whether the system meets the

overall safety assurance goals. There is a feedback loop through the *query and concretize* part in case additional tests are required to make a well-informed decision.

The *coverage* analysis provides quantitative assessments of the extent to which the scenario set addresses relevant operational conditions. Therefore, the coverage metrics can be directly used for this. In addition, metrics related to the dissimilarity of scenarios support the evaluation of scenario set diversity, ensuring that the testing scenarios are diverse.

The *test evaluate* part assesses each test execution to determine whether the test has been executed well and to interpret the outcome of individual tests. This part mainly uses the information provided from the test cases and the test objectives, so no further scenario metrics are directly involved.

The *safety case* compiles structured, evidence-based arguments to demonstrate that an ADS meets (legal) safety standards and is ready for deployment. Coverage metrics help to justify that the system has been validated across all relevant operational contexts. Exposure metrics can serve as weighting factors in the safety argument, enhancing the credibility of risk-based safety assessments.

The *decide* block finalizes the safety assurance process by integrating results from earlier steps into a binary pass/fail decision. This part relies on inputs from prior components to support a traceable and auditable outcome aligned with regulatory expectations, so no further scenario metrics are directly involved.

4.4. SAF Component—ISMR

Upon a positive decision following the *safety argument* component, the ADS may be deployed with ISMR in place. The ISMR serves multiple purposes [4]. First, it monitors the system during deployment, thereby ensuring continuous safety. To measure continuous safety, scenario criticality metrics might be used. Second, ISMR enables the continuous collection of evidence supporting the assumptions used during the safety case. For example, assumptions on the exposure of scenarios—possibly based on scenario statistics from the SCDBs—can be verified during deployment. In this way, ISMR provides additional input to the *safety argument* component. Third, new scenarios may be detected, e.g., with the use of dissimilarity metrics, and may be included into the SCDBs. As a result, ISMR serves as one of the inputs for the *scenario* component.

4.5. SAF Component—Audit

The *audit* component evaluates the manufacturer's safety management processes, including how they identify, analyze, and mitigate risks throughout the development and deployment of the ADS. This goes beyond just passing specific technical tests; the *audit* ensures that the manufacturer adopts a structured, transparent, and accountable approach to safety. A key aspect of this process is the use of the SCDBs to derive test scenarios that adequately cover the system's ODD. In this context, coverage metrics can support the *audit*.

5. Discussion

This article has provided a structured overview of metrics relevant to scenario-based assessments of ADSs within the context of the SUNRISE SAF. This section discusses several limitations and remaining challenges.

First, many of the discussed metrics are inherently use-case-dependent. For example, scenario exposure metrics depend on the actual ODD of an ADS. Similarly, completeness metrics, though conceptually transferable, must be tailored to the operational context of the ADS under test. In those cases, it is not particularly useful to already add those metrics as metadata to the scenarios in the SCDBs at the *store* phase of the SAF. Given these dependencies, it is difficult to make general statements on aspects of a particular

SCDB without providing some context. For example, if a particular ADS does not respond differently to vehicles of different colors (e.g., because it does not depend on camera footage) and if a vehicle's color is omitted from the scenario description, the scenario description might still be considered complete for testing this particular ADS. However, for an ADS that may show different behavior based on a vehicle's color, the scenario description must contain the vehicle's color in order to be complete.

Second, for different types of metrics, there is generally no single, universal metric that fully captures the concept. There are different aspects within the assessment, thus requiring different metrics. For example, completeness can refer to the inclusion of all relevant scenario parameters, the specification of boundary conditions, or the presence of required environmental elements; each of these may require distinct measurement approaches. Similarly, coverage can pertain to coverage of parameter ranges, environmental conditions within the ODD, or behavioral variations in traffic participants, etc. This highlights the need for multiple interpretable metrics that can be combined or adapted depending on the evaluation objectives, rather than relying on a one-size-fits-all solution.

While metric values can provide useful insights, they often require further interpretation to fully understand the characteristics of a scenario set. For example, a low coverage score might suggest that important scenarios are not addressed, potentially due to gaps in scenario generation or selection. However, it could also reflect the absence of scenarios that are unlikely or irrelevant within the defined ODD, such as rainy weather conditions inside a tunnel. In such cases, additional analysis may be needed to determine whether low coverage truly indicates a deficiency or simply reflects the operational reality. This underscores the importance of contextualizing metric results rather than relying on absolute values alone.

That said, metrics play a critical role in enabling iterative refinement across SAF stages. For example, coverage and criticality metrics can be used to identify underrepresented or safety-critical regions in the scenario space, which in turn inform new scenario instantiations. However, managing these feedback loops effectively remains a challenge, especially when metrics are applied at multiple levels of abstraction (e.g., logical vs. concrete scenarios) and across diverse testing environments (e.g., virtual simulation, proving ground, and hardware-in-the-loop).

While this work focused on reviewing existing scenario metrics and linking them to SAF components, interactions and trade-offs between metrics deserve further analysis. For example, increasing scenario diversity may also reduce representativeness if rare edge cases dominate. Similarly, criticality and exposure can pull in opposite directions, as highly critical scenarios are often low-frequency. Understanding such interdependencies and developing methods for balancing or aggregating metrics will be crucial to make SAF-based safety arguments both robust and efficient. We consider this an important direction for future research.

An additional avenue for future research lies in linking scenario metrics with currently trending developments, particularly large language models. Recent surveys (e.g., [8]) highlight how these models can support scenario-based testing by automatically generating diverse, realistic, and critical scenarios, as well as by assisting in the interpretation of test outcomes. These capabilities complement scenario metrics: for instance, scenarios generated using large language models could be evaluated using completeness, coverage, and exposure metrics to ensure quality and relevance. Exploring how such artificial-intelligence-driven methods can be integrated with structured metric frameworks like the SAF represents a promising direction for enhancing both the scalability and the explanatory power of scenario-based safety assurance.

6. Conclusions

This paper has provided a structured overview of scenario metrics that support scenario-based safety assessments of ADS within the SUNRISE SAF. We have identified six core categories of metrics: scenario completeness, scenario coverage, scenario criticality, scenario diversity/dissimilarity, scenario exposure, and scenario complexity. In addition, other related metrics, such as realism, rarity, and representativeness, have been briefly discussed. The relevance of the six main metric types to the SAF stages, such as scenario generation, scenario allocation, test execution, and coverage analysis, have been analyzed. The presented metrics play a foundational role in developing the safety case of an ADS, which ultimately enables the deployment of these systems on public roads.

A key insight from this work is that there is no one-size-fits-all metric for any category; different facets of each concept may need to be captured using multiple, context-dependent metrics. Furthermore, many metrics are interdependent or overlapping, requiring careful coordination in their application to avoid redundancy or misinterpretation. In addition, rather than relying on absolute values alone, it remains important to contextualize metric values. As ADS assessment practices evolve, there is a need for further formalization of metrics, validation through real-world data, and tooling support to integrate these metrics effectively within test and validation pipelines.

Future research should focus on defining and validating emerging metrics, developing aggregation strategies that reflect both system performance and risk, and supporting regulators and developers in interpreting metric values within diverse operational and regulatory contexts. Future work should also address the interactions and trade-offs between metrics, for example, balancing coverage against representativeness or exposure against criticality, in order to provide more nuanced and actionable guidance for applying the SAF. By advancing the metric landscape in a structured and scenario-centric way, we move closer to scalable, explainable, and robust safety assurance for ADSs.

Author Contributions: Conceptualization, E.d.G., T.S., F.H.-S. and S.V.B.; methodology, E.d.G., T.S., F.H.-S. and S.V.B.; writing—original draft preparation, E.d.G., T.S., F.H.-S. and S.V.B.; writing—review and editing, E.d.G. and O.O.d.C.; visualization, E.d.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in this work has been made possible by the SUNRISE project. This project is funded by the European Union's Horizon Europe Research and Innovation Actions under grant agreement No. 101069573. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Kalra, N.; Paddock, S.M. Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability? *Transp. Res. Part A Policy Pract.* **2016**, *94*, 182–193. [CrossRef]
2. 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]
3. de Gelder, E.; Op den Camp, O.; Broos, J.; Paardekooper, J.P.; van Montfort, S.; Kalisvaart, S.; Goossens, H. *Scenario-Based Safety Assessment of Automated Driving Systems*; Technical Report; TNO: Helmond, The Netherlands, 2024.
4. Op den Camp, O.; de Gelder, E. Operationalization of Scenario-Based Safety Assessment of Automated Driving Systems. In Proceedings of the IEEE International Automated Vehicle Validation Conference, Baden-Baden, Germany, 30 September–2 October 2025. [CrossRef]

5. ECE/TRANS/WP.29/2021/61. *New Assessment/Test Method for Automated Driving (NATM)–Master Document*; Technical Report; World Forum for Harmonization of Vehicle Regulations: Geneva, Switzerland, 2021.
6. Finkeldei, F.; Thees, C.; Weghorn, J.N.; Althoff, M. Scenario Factory 2.0: Scenario-Based Testing of Automated Vehicles with CommonRoad. *Automot. Innov.* **2025**, *8*, 207–220. [CrossRef]
7. Althoff, M.; Koschi, M.; Manzinger, S. CommonRoad: Composable Benchmarks for Motion Planning on Roads. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; Volume 6, pp. 719–726. [CrossRef]
8. Zhao, Y.; Zhou, J.; Bi, D.; Mihalj, T.; Hu, J.; Eichberger, A. A Survey on the Application of Large Language Models in Scenario-Based Testing of Automated Driving Systems. *arXiv* **2025**, arXiv:2505.16587. [CrossRef]
9. Yan, S.; Zhang, X.; Hao, K.; Xin, H.; Luo, Y.; Yang, J.; Fan, M.; Yang, C.; Sun, J.; Yang, Z. On-Demand Scenario Generation for Testing Automated Driving Systems. *ACM Softw. Eng.* **2025**, *2*, 86–105. [CrossRef]
10. Emran, N.A. Data Completeness Measures. In *Pattern Analysis, Intelligent Security and the Internet of Things*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 117–130. [CrossRef]
11. Westhofen, L.; Neurohr, C.; Koopmann, T.; Butz, M.; Schütt, B.; Utesch, F.; Neurohr, B.; Gutenkunst, C.; Böde, E. Criticality Metrics for Automated Driving: A Review and Suitability Analysis of the State of the Art. *Arch. Comput. Methods Eng.* **2022**, *30*, 1–35. [CrossRef]
12. Wang, C.; Xie, Y.; Huang, H.; Liu, P. A Review of Surrogate Safety Measures and Their Applications in Connected and Automated Vehicles Safety Modeling. *Accid. Anal. Prev.* **2021**, *157*, 106157. [CrossRef]
13. Liu, T.; Wang, C.; Yin, Z.; Mi, Z.; Xiong, X.; Guo, B. Complexity Quantification of Driving Scenarios with Dynamic Evolution Characteristics. *Entropy* **2024**, *26*, 1033. [CrossRef]
14. Tahir, Z.; Alexander, R. Coverage Based Testing for V&V and Safety Assurance of Self-Driving Autonomous Vehicles: A Systematic Literature Review. In Proceedings of the IEEE International Conference On Artificial Intelligence Testing (AITest), Oxford, UK, 3–6 August 2020; pp. 23–30. [CrossRef]
15. *ISO 34502*; Road Vehicles–Test Scenarios for Automated Driving Systems–Engineering Framework and Process of Scenario-Based Safety Evaluation. International Organization for Standardization: Geneva, Switzerland, 2022.
16. Glasmacher, C.; Weber, H.; Eckstein, L. Towards a Completeness Argumentation for Scenario Concepts. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Jeju Island, Republic of Korea, 2–5 June 2024. [CrossRef]
17. de Gelder, E.; Buermann, M.; Op den Camp, O. Coverage Metrics for a Scenario Database for the Scenario-Based Assessment of Automated Driving Systems. In Proceedings of the IEEE International Automated Vehicle Validation Conference, Pittsburgh, PA, USA, 21–23 October 2024. [CrossRef]
18. Cai, J.; Deng, W.; Guang, H.; Wang, Y.; Li, J.; Ding, J. A Survey on Data-Driven Scenario Generation for Automated Vehicle Testing. *Machines* **2022**, *10*, 1101. [CrossRef]
19. Jansson, J. Collision Avoidance Theory: With Application to Automotive Collision Mitigation. Ph.D., Thesis, Linköping University Electronic Press, Linköping, Sweden, 2005.
20. Hayward, J.C. *Near Miss Determination Through Use of a Scale of Danger*; Technical Report TTSC-7115; Pennsylvania State University: University Park, PA, USA, 1972.
21. Wachenfeld, W.; Junietz, P.; Wenzel, R.; Winner, H. The Worst-Time-to-Collision Metric for Situation Identification. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Gotenburg, Sweden, 19–22 June 2016; pp. 729–734. [CrossRef]
22. Eggert, J. Predictive Risk Estimation for Intelligent ADAS Functions. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 711–718. [CrossRef]
23. Minderhoud, M.M.; Bovy, P.H. Extended Time-to-Collision Measures for Road Traffic Safety Assessment. *Accid. Anal. Prev.* **2001**, *33*, 89–97. [CrossRef]
24. Varhelyi, A. Drivers' Speed Behaviour at a Zebra Crossing: A Case Study. *Accid. Anal. Prev.* **1998**, *30*, 731–743. [CrossRef] [PubMed]
25. Allen, B.L.; Shin, B.T.; Cooper, P.J. Analysis of Traffic Conflicts and Collisions. *Transp. Res. Board* **1978**, *667*, 67–74.
26. Brannstrom, M.; Sjoberg, J.; Coelingh, E. A Situation and Threat Assessment Algorithm for a Rear-End Collision Avoidance System. In Proceedings of the IEEE Intelligent Vehicles Symposium, Eindhoven, The Netherlands, 4–6 June 2008; pp. 102–107. [CrossRef]
27. Yue, B.; Shi, S.; Wang, S.; Lin, N. Low-Cost Urban Test Scenario Generation Using Microscopic Traffic Simulation. *IEEE Access* **2020**, *8*, 123398–123407. [CrossRef]
28. Cafiso, S.; Garcia, A.G.; Cavarra, R.; Rojas, M.R. Crosswalk Safety Evaluation Using a Pedestrian Risk Index as Traffic Conflict Measure. In Proceedings of the 3rd International Conference on Road safety and Simulation, Indianapolis, IN, USA, 14–16 September 2011; Volume 15.
29. Cunto, F.; Saccomanno, F.F. Calibration and Validation of Simulated Vehicle Safety Performance at Signalized Intersections. *Accid. Anal. Prev.* **2008**, *40*, 1171–1179. [CrossRef] [PubMed]

30. Schreier, M.; Willert, V.; Adamy, J. An integrated approach to maneuver-based trajectory prediction and criticality assessment in arbitrary road environments. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2751–2766. [CrossRef]
31. Huber, B.; Herzog, S.; Sippl, C.; German, R.; Djanatliev, A. Evaluation of Virtual Traffic Situations for Testing Automated Driving Functions Based on Multidimensional Criticality Analysis. In Proceedings of the IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Virtual, 20–23 September 2020; pp. 1–7. [CrossRef]
32. Baumann, D.; Pfeffer, R.; Sax, E. Automatic Generation of Critical Test Cases for the Development of Highly Automated Driving Functions. In Proceedings of the IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Helsinki, Finland, 25 April–19 May 2021; pp. 1–5. [CrossRef]
33. Piziali, A. *Functional Verification Coverage Measurement and Analysis*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007. [CrossRef]
34. Alexander, R.; Hawkins, H.; Rae, D. *Situation Coverage—A Coverage Criterion for Testing Autonomous Robots*; Technical Report; Department of Computer Science, University of York: York, UK, 2015.
35. Araujo, H.; Mousavi, M.R.; Varshosaz, M. Testing, Validation, and Verification of Robotic and Autonomous Systems: A Systematic Review. *ACM Trans. Softw. Eng. Methodol.* **2023**, *32*, 1–61. [CrossRef]
36. de Gelder, E.; Paardekooper, J.P.; Op den Camp, O.; De Schutter, B. Safety Assessment of Automated Vehicles: How to Determine Whether We Have Collected Enough Field Data? *Traffic Inj. Prev.* **2019**, *20*, 162–170. [CrossRef]
37. Hartjen, L.; Philipp, R.; Schuldt, F.; Friedrich, B. Saturation effects in recorded maneuver data for the test of automated driving. In Proceedings of the 13. Uni-DAS eV Workshop Fahrerassistenz und Automatisiertes Fahren, 16–17 July 2020; pp. 74–83.
38. Glasmacher, C.; Schuldes, M.; Weber, H.; Wagener, N.; Eckstein, L. Acquire Driving Scenarios Efficiently: A Framework for Prospective Assessment of Cost-Optimal Scenario Acquisition. In Proceedings of the IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; pp. 1971–1976. [CrossRef]
39. Tatar, M. Enhancing ADAS Test and Validation with Automated Search for Critical Situations. In Proceedings of the Driving Simulation Conference (DSC), Tübingen, Germany, 16–18 September 2015; pp. 1–4.
40. Mullins, G.E.; Stankiewicz, P.G.; Hawthorne, R.C.; Gupta, S.K. Adaptive Generation of Challenging Scenarios for Testing and Evaluation of Autonomous Vehicles. *J. Syst. Softw.* **2018**, *137*, 197–215. [CrossRef]
41. Alnaser, A.J.; Sargolzaei, A.; Akbaş, M.I. Autonomous Vehicles Scenario Testing Framework and Model of Computation: On Generation and Coverage. *IEEE Access* **2021**, *9*, 60617–60628. [CrossRef]
42. 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]
43. Tian, Y.; Pei, K.; Jana, S.; Ray, B. DeepTest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars. In Proceedings of the 40th International Conference on Software Engineering, Gothenburg, Sweden, 27 May–3 June 2018; pp. 303–314. [CrossRef]
44. Laurent, T.; Klikovits, S.; Arcaini, P.; Ishikawa, F.; Ventresque, A. Parameter Coverage for Testing of Autonomous Driving Systems under Uncertainty. *ACM Trans. Softw. Eng. Methodol.* **2023**, *32*, 1–31. [CrossRef]
45. Zhu, B.; Zhang, P.; Zhao, J.; Deng, W. Hazardous Scenario Enhanced Generation for Automated Vehicle Testing Based on Optimization Searching Method. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 7321–7331. [CrossRef]
46. Zhong, Z.; Kaiser, G.; Ray, B. Neural Network Guided Evolutionary Fuzzing for Finding Traffic Violations of Autonomous Vehicles. *IEEE Trans. Softw. Eng.* **2022**, *49*, 1860–1875. [CrossRef]
47. Ries, L.; Rigoll, P.; Braun, T.; Schulik, T.; Daube, J.; Sax, E. Trajectory-Based Clustering of Real-World Urban Driving Sequences with Multiple Traffic Objects. In Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 1251–1258. [CrossRef]
48. Lin, Q.; Wang, W.; Zhang, Y.; Dolan, J.M. Measuring Similarity of Interactive Driving Behaviors Using Matrix Profile. In Proceedings of the American Control Conference (ACC), Denver, CO, USA, 1–3 July 2020; pp. 3965–3970. [CrossRef]
49. Kerber, J.; Wagner, S.; Groh, K.; Notz, D.; Kühbeck, T.; Watzenig, D.; Knoll, A. Clustering of the Scenario Space for the Assessment of Automated Driving. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 578–583. [CrossRef]
50. Kruber, F.; Wurst, J.; Botsch, M. An Unsupervised Random Forest Clustering Technique for Automatic Traffic Scenario Categorization. In Proceedings of the 21st International conference on intelligent transportation systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 2811–2818. [CrossRef]
51. Nguyen, V.; Huber, S.; Gambi, A. SALVO: Automated Generation of Diversified Tests for Self-Driving Cars from Existing Maps. In Proceedings of the IEEE International Conference on Artificial Intelligence Testing (AITest), Oxford, UK, 23–26 August 2021; pp. 128–135. [CrossRef]
52. Zohdinasab, T.; Riccio, V.; Gambi, A.; Tonella, P. Deephyperion: Exploring the Feature Space of Deep Learning-Based Systems through Illumination Search. In Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis, Virtual, 11–17 July 2021; pp. 79–90. [CrossRef]

53. Wheeler, T.A.; Kochenderfer, M.J. Critical Factor Graph Situation Clusters for Accelerated Automotive Safety Validation. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 2133–2139. [CrossRef]
54. Mahadikar, B.B.; Rajesh, N.; Kurian, K.T.; Lefeber, E.; Ploeg, J.; van de Wouw, N.; Alirezaei, M. Formulating a dissimilarity metric for comparison of driving scenarios for Automated Driving Systems. In Proceedings of the 2024 IEEE Intelligent Vehicles Symposium (IV), Jeju Island, Republic of Korea, 2–5 June 2024; pp. 1091–1098. [CrossRef]
55. Dokania, N.; Singh, T.; Lefeber, E.; Ploeg, J.; Alirezaei, M. Implementing a dissimilarity metric for scenarios categorization and selection for automated driving systems. *IFAC-PapersOnLine* **2025**, *59*, 145–150. [CrossRef]
56. Tian, H.; Jiang, Y.; Wu, G.; Yan, J.; Wei, J.; Chen, W.; Li, S.; Ye, D. MOSAT: Finding Safety Violations of Autonomous Driving Systems Using Multi-Objective Genetic Algorithm. In Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Singapore, 14–16 November 2022; pp. 94–106. [CrossRef]
57. de Gelder, E.; Elrofai, H.; Khabbaz Saberi, A.; Op den Camp, O.; Paardekooper, J.P.; De Schutter, B. Risk Quantification for Automated Driving Systems in Real-World Driving Scenarios. *IEEE Access* **2021**, *9*, 168953–168970. [CrossRef]
58. Hakkert, A.S.; Braimaister, L.; Van Schagen, I. *The Uses of Exposure and Risk in Road Safety Studies*; Technical Report R-2002-12; SWOV Institute for Road Safety: The Hague, The Netherlands, 2002.
59. Gietelink, O. Design and Validation of Advanced Driver Assistance Systems. Ph.D. Thesis, Delft University of Technology, Delft, The Netherlands, 2007.
60. de Gelder, E.; Op den Camp, O. How Certain Are We That Our Automated Driving System Is Safe? *Traffic Inj. Prev.* **2023**, *24*, S131–S140. [CrossRef]
61. Feng, S.; Yan, X.; Sun, H.; Feng, Y.; Liu, H.X. Intelligent Driving Intelligence Test for Autonomous Vehicles with Naturalistic and Adversarial Environment. *Nat. Commun.* **2021**, *12*, 748. [CrossRef]
62. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
63. Chen, Y.C. A Tutorial on Kernel Density Estimation and Recent Advances. *Biostat. Epidemiol.* **2017**, *1*, 161–187. [CrossRef]
64. *E/ECE/TRANS/505/Rev.3/Add.156*; Uniform Provisions Concerning the Approval of Vehicles with Regard to Automated Lane Keeping Systems. World Forum for Harmonization of Vehicle Regulations: Geneva, Switzerland, 2021.
65. Nakamura, H.; Muslim, H.; Kato, R.; Préfontaine-Watanabe, S.; Nakamura, H.; Kaneko, H.; Imanaga, H.; Antona-Makoshi, J.; Kitajima, S.; Uchida, N.; et al. Defining Reasonably Foreseeable Parameter Ranges Using Real-World Traffic Data for Scenario-Based Safety Assessment of Automated Vehicles. *IEEE Access* **2022**, *10*, 37743–37760. [CrossRef]
66. Muslim, H.; Endo, S.; Imanaga, H.; Kitajima, S.; Uchida, N.; Kitahara, E.; Ozawa, K.; Sato, H.; Nakamura, H. Cut-out Scenario Generation with Reasonability Foreseeable Parameter Range from Real Highway Dataset for Autonomous Vehicle Assessment. *IEEE Access* **2023**, *11*, 45349–45363. [CrossRef]
67. de Gelder, E.; Op den Camp, O. A Quantitative Method to Determine What Collisions Are Reasonably Foreseeable and Preventable. *Saf. Sci.* **2023**, *167*, 106233. [CrossRef]
68. Sussman, J.M. *Perspectives on Intelligent Transportation Systems (ITS)*; Springer: Berlin/Heidelberg, Germany, 2005.
69. Issler, M.; Goss, Q.; Akbaş, M.İ. Complexity Evaluation of Test Scenarios for Autonomous Vehicle Safety Validation Using Information Theory. *Information* **2024**, *15*, 772. [CrossRef]
70. Yu, R.; Zheng, Y.; Qu, X. Dynamic driving environment complexity quantification method and its verification. *Transp. Res. Part C Emerg. Technol.* **2021**, *127*, 103051. [CrossRef]
71. Dunne, R.; Schatz, S.; Fiore, S.M.; Martin, G.; Nicholson, D. Scenario-Based Ttraining: Scenario Complexity. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, San Francisco, CA, USA, 27 September–1 October 2010; pp. 2238–2242. [CrossRef]
72. Faure, V.; Lobjois, R.; Benguigui, N. The Effects of Driving Environment Complexity and Dual Tasking on Drivers' Mental Workload and Eye Blink Behavior. *Transp. Res. Part F TRaffic Psychol. Behav.* **2016**, *40*, 78–90. [CrossRef]
73. Manawadu, U.E.; Kawano, T.; Murata, S.; Kamezaki, M.; Sugano, S. Estimating Driver Workload with Systematically Varying Traffic Complexity Using Machine Learning: Experimental Design. In Proceedings of the International Conference on Intelligent Human Systems Integration, Dubai, United Arab Emirates, 7–9 January 2018; pp. 106–111. [CrossRef]
74. Liu, Y.; Hansen, J.H. Towards Complexity Level Classification of Driving Scenarios Using Environmental Information. In Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 810–815. [CrossRef]
75. Berseth, G.; Kapadia, M.; Faloutsos, P. SteerPlex: Estimating Scenario Complexity for Simulated Crowds. In Proceedings of the Motion on Games, Dublin, UK, 6–8 November 2013; pp. 67–76. [CrossRef]
76. Wang, J.; Zhang, C.; Liu, Y.; Zhang, Q. Traffic Sensory Data Classification by Quantifying Scenario Complexity. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Suzhou, China, 26–30 June 2018; pp. 1543–1548. [CrossRef]

77. Zhang, L.; Ma, Y.; Xing, X.; Xiong, L.; Chen, J. Research on the Complexity Quantification Method of Driving Scenarios Based on Information Entropy. In Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 3476–3481. [CrossRef]
78. Zhou, J.; Wang, L.; Wang, X. Online adaptive generation of critical boundary scenarios for evaluation of autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 6372–6388. [CrossRef]
79. Association for Standardization of Automation and Measuring Systems. ASAM OpenSCENARIO XML, 2024. Available online: https://www.asam.net/standards/detail/openscenario (accessed on 12 September 2025).
80. *ISO 34504*; Road Vehicles–Test Scenarios for Automated Driving Systems–Scenario Categorization. International Organization for Standardization: Geneva, Switzerland, 2024.
81. Neurohr, C.; Westhofen, L.; Butz, M.; Bollmann, M.; Eberle, U.; Galbas, R. Criticality Analysis for the Verification and Validation of Automated Vehicles. *IEEE Access* **2021**, *9*, 18016–18041. [CrossRef]
82. de Gelder, E.; Paardekooper, J.P.; Khabbaz Saberi, A.; Elrofai, H.; Op den Camp, O.; Kraines, S.; Ploeg, J.; De Schutter, B. Towards an Ontology for Scenario Definition for the Assessment of Automated Vehicles: An Object-Oriented Framework. *IEEE Trans. Intell. Veh.* **2022**, *7*, 300–314. [CrossRef]
83. *ISO 26262*; Road Vehicles–Functional Safety. International Organization for Standardization: Geneva, Switzerland, 2018.
84. Paardekooper, J.P.; Montfort, S.; Manders, J.; Goos, J.; de Gelder, E.; Op den Camp, O.; Bracquemond, A.; Thiolon, G. Automatic Identification of Critical Scenarios in a Public Dataset of 6000 km of Public-Road Driving. In Proceedings of the 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV), Eindhoven, The Netherlands, 10–13 June 2019.
85. Sagmeister, S.; Kounatidis, P.; Goblirsch, S.; Lienkamp, M. Analyzing the Impact of Simulation Fidelity on the Evaluation of Autonomous Driving Motion Control. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Jeju Island, Republic of Korea, 2–5 June 2024; pp. 230–237. [CrossRef]
86. Akella, P.; Ubellacker, W.; Ames, A.D. Safety-Critical Controller Verification via Sim2Real Gap Quantification. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 10539–10545. [CrossRef]
87. Sangeerth, P.; Jagtap, P. Quantification of Sim2Real Gap via Neural Simulation Gap Function. *arXiv* **2025**, arXiv:2506.17675. [CrossRef]
88. Park, S.; Pahk, J.; Jahn, L.L.F.; Lim, Y.; An, J.; Choi, G. A Study on Quantifying Sim2real Image Gap in Autonomous Driving Simulations Using Lane Segmentation Attention Map Similarity. In Proceedings of the International Conference on Intelligent Autonomous Systems, Qinhuangdao, China, 22–24 September 2023; pp. 203–212. [CrossRef]
89. Mahajan, I.; Unjhawala, H.; Zhang, H.; Zhou, Z.; Young, A.; Ruiz, A.; Caldararu, S.; Batagoda, N.; Ashokkumar, S.; Negrut, D. Quantifying the Sim2real gap for GPS and IMU sensors. *arXiv* **2024**, arXiv:2403.11000. [CrossRef]
90. Pahk, J.; Shim, J.; Baek, M.; Lim, Y.; Choi, G. Effects of Sim2Real Image Translation via DCLGAN on Lane Keeping Assist System in CARLA Simulator. *IEEE Access* **2023**, *11*, 33915–33927. [CrossRef]
91. Waheed, A.; Areti, M.; Gallantree, L.; Hasnain, Z. Quantifying the Sim2Real Gap: Model-Based Verification and Validation in Autonomous Ground Systems. *IEEE Robot. Autom. Lett.* **2025**, *10*, 3819–3826. [CrossRef]
92. Petrucelli, E.; States, J.D.; Hames, L.N. The Abbreviated Injury Scale: Evolution, Usage and Future Adaptability. *Accid. Anal. Prev.* **1981**, *13*, 29–35. [CrossRef]
93. de Gelder, E.; Hof, J.; Cator, E.; Paardekooper, J.P.; Op den Camp, O.; Ploeg, J.; De Schutter, B. Scenario Parameter Generation Method and Scenario Representativeness Metric for Scenario-Based Assessment of Automated Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 18794–18807. [CrossRef]
94. Zhang, J.X.; Op den Camp, O.; de Vries, S.; Bourauel, B.; Hillbrand, B.; Nieto Doncel, M.; Gronvall, J.F.; Stern, D.; Bolovinou, A.; Arrieta Fernández, A.; et al. *SUNRISE D2.3: Final SUNRISE Safety Assurance Framework*; Technical Report; European Union: 2025.
95. Scholtes, M.; Schuldes, M.; Weber, H.; Wagener, N.; Hoss, M.; Eckstein, L. OMEGAFormat: A comprehensive format of traffic recordings for scenario extraction. In Proceedings of the Workshop Fahrerassistenz und automatisiertes Fahren, Berkheim, Germany, 9–11 May 2022; pp. 195–205.
96. Beckmann, J.; Torres Camara, J.M.; Kaynar, E.; de Gelder, E.; Daskalaki, E.; Hillbrand, B.; Amler, T.; Thorsén, A.; Irvine, P.; Kirchengast, M.; et al. *SUNRISE D3.4: Report on Subspace Creation Methodology*; Technical Report; European Union: 2025; *in preparation*.
97. Hillbrand, B.; Kirchengast, M.; Ballis, A.; Panagiotopoulos, I.; Menzel, T.; Amler, T.; Collado, E.; Beckmann, J.; Skoglund, M.; Thorsén, A.; et al. *SUNRISE D3.3: Report on the Initial Allocation of Scenarios to Test Instances*; Technical Report; European Union: 2024.