International Alliance for Mobility Testing and Standardization™ Practice Report

IAMTS0002202301

Issued 2023-01-20

FOR COMMENT

Worldwide Comparison of Highly Automated Vehicle Testing Scenarios Using Real Road Statistics

Rationale

Compared to SAE L1 and L2 vehicles on the road today (e.g., adaptive cruise control, lane-keeping assist), Highly Automated Vehicles (HAVs) need to be comprehensively validated through a large scale and wide-range of tests consisting of various combinations of elements such as the road environment, traffic participants, environmental conditions and more. It is vital to ensure the public trust these HAVs, and critical for all vehicle OEMs and operators to openly demonstrate the capabilities of their products to the world. Around the world different value chain stakeholders have tackled this problem by using scenario-based testing.

The idea that a HAV could pass all existing scenarios with flying colors can help to build confidence within this industry. Some of the HAV testing has been done on the public road or real-world testing, whereas some are virtual tests or tested in a simulation environment. All of these test cases should share the same language: both in terms of the description and in the procedure.

With the help of different value chain stakeholders and a specific comparison process, a global community of experts shared the same vision of the need to harmonize on the testing scenarios and test cases, focusing on applying real world statistics, and demonstrating how scenario-based testing is used within this community and how it can benefit all stakeholders worldwide.

Preface

IAMTS is a global, membership-based association of organizations that are stakeholders in the testing, standardization, and certification of advanced mobility systems and services. IAMTS brings together testing consumers and providers at a global scale to help develop a commonly accepted framework of test scenarios, validation and certification methods, and terminology.

Our mission is to develop and grow an international portfolio of advanced mobility testbeds that meet the highest quality implementation and operational standards.

Our vision is to create a global community of advanced mobility testing service providers with companies, organizations, and agencies in need of such services; to learn, develop, and share best practices to ensure consistent, replicable, and reliable testing; to maintain a global directory of physical, virtual, and cyber-physical testbeds and support and promote their audited capabilities; and to promote the rapid evolution of standards and certifications to ensure the safe deployment of advanced mobility systems and services.

Introduction

The purpose of this paper is to illustrate a concept for generating test cases for an internationally harmonized verification and validation (V&V) procedure for highly automated vehicles (HAVs). In this paper, HAVs refer to vehicles that operate with Level 4 automation features as defined by the Society of Automotive Engineers (SAE) [1]. A simpler way to understand them is that they are vehicles chauffeured by a robotic driver but have a limited operational design domain (ODD). HAV developers predominantly use three modes for their internal V&V: simulations, testing inside closed test tracks or lab testing, and drives on the public roads. As a complex system of new technologies held together by intelligent software, HAVs need to undergo extensive testing. Simulation testing is often considered an option to reduce the cost and time of testing. However, due to the challenges in model calibration, simulation is mainly used in the research and development (R&D) process of HAVs, focusing on checking the decision algorithms of Automated Driving Systems (ADS). Whereas the perception system, control system, human-machine interaction system, auxiliary environment system and vehicle interoperability are still largely tested in real environments with real vehicles. Real-world testing on public roads still faces many challenges to overcome, including time and cost of road testing, regional laws and regulations, and public safety. Lab testing falls somewhere between virtual simulations and public road testing, which will likely be a primary tool for V&V overseen by public agencies.

HAVs have the potential to improve transportation safety, convenience and accessibility for people and goods, but a key challenge today remains that many people do not trust or want to use them. HAVs are not available for purchase from any automotive OEMs anywhere in the world yet. However, there are autonomous people movers being offered by pilot fleets in the US, Europe, China and Japan in geofenced zones and on fixed routes. In addition, goods deliveries using HAVs are also available in many parts of the world. These pilot deployments were possible largely through ad-hoc permissions or exemptions on a case-by-case basis. A critical step before HAVs can scale-up to become successful business operations is a rigorous and trustworthy V&V process. A global community of experts from various fields including HAV researchers, developers and V&V practitioners from Europe, Asia and the United States have collaborated to present an early attempt to build the concept for a globally harmonized test procedure. These techniques are not intended to be a static tome, but rather a spark for collaborative discussion, trial, and further refinement.

As a global alliance many expert practitioners from IAMTS members spanning different regions, different value chain stakeholders (testing, simulation & certification, proving ground, toolchain provider, applied research) have provided their best practices in dealing with a global acceptance scenario data comparison and usage. The value of these different perspectives and practices led to an insightful diversity of views on what is expected from scenarios / test cases, how to structure them in databases, how to exploit these, and how to interpret their usefulness. The engagement also led to bringing together useful and proven methods practiced for different workflows, but also resulting commonalities and divergence of views, which will help us to derive requirements on how to use test cases, scenarios, and databases for the future.

In memory of Huei Peng for his relentless dedication to engineering autonomy, to his research at Mcity, and his time spent contributing to this paper.

Table of Contents

1. Challenges to Testing Highly Automated Vehicles Across the World

Compared to SAE L1 and L2 vehicles on the road today (e.g., adaptive cruise control, lane-keeping assist), HAVs need to be comprehensively validated through a large scale and wide-range of tests consisting of various combinations of elements such as road environment, traffic participants, environmental conditions and more. The purpose of this paper is to outline our efforts in compiling knowledge from global researchers and present learnings toward the development of an internationally harmonized scenario-based assessment, which is an essential step in the development of new vehicle certification procedures for HAVs. Scenario-based assessment typically considers the complete vehicle, and such scenario-based assessment is an addition to testing for Functional Safety (ISO 26262), and Safety of the Intended Functionality (ISO/PAS 21448).

Existing testing frameworks are based on a limited number of test matrices that are no longer sufficient to ensure operational safety of HAVs on the road. A new framework is needed, which includes scenario-based assessment, relying on extensive simulations, and executing cyber-physical testing. To provide adequate test coverage, HAV researchers usually start by collecting petabytes of data using highly instrumented vehicles. Ideally, the collected data and their characteristics and probabilities of occurrence (exposure) would densely cover the entire range of real-world traffic situations that might be encountered by the HAVs. However, no single entity can realistically compile sufficiently extensive amount of data on every single scenario possible. Furthermore, data from different regions in the world may show unique characteristics that may not be realized without real world testing data. IAMTS believes comparing and analyzing data collected from different parts of the world is an important first step to understanding what challenges need to be addressed to safely operate and deploy highly automated vehicles around the world.

The collaborating practitioners brought together experience and real-world highway data sets from three regions: Europe (Netherlands), USA (Great Lakes Region), and China (Jing Jin Ji Metropolitan Area). The nature of the data is therefore quite diverse in road network, population density, and operational domains. The collaborators also used the data for different purposes like applying methodologies to extract different insights on the validation of HAVs. The benefits of utilizing a consortium model to bring this data together in a pre-competitive and neutral manner allows all stakeholders to participate in the discussion and creation of best practices and guidelines.

In this paper, we will independently be looking into the methodology from each region and will eventually converge into a robust proposal with identified constraints. The first major constraint would be the willingness and the approval of the local governments, this is especially the case for public road data collection and testing. This is because data collection and testing on public roads is time-consuming, costly, and potentially unsafe. The lack of clear regulations and insurance claims mechanisms for HAVs has limited the development of testing on public roads. As HAVs are not yet proven to be operationally safe under all conditions, Europe and Asia largely have maintained a cautious attitude towards opening public road testing for HAVs, relying only on part of the public roads with selected special zones, which limit real-world exposure. In addition, the perception of authorities making exemptions for HAVs that are not yet proven to be operationally safe is concerning. Data collection for building a scenario database for HAV testing is done on public roads on a regular basis and in cooperation with HAV industry, as it does not require the automated driving functionality to be active and consequently does not pose a safety risk.

US auto regulators have stayed with the tradition of a laissez-faire approach to vehicle testing and certification. However, letting the government or a third party access the simulation stack of the HAV developers causes serious confidentiality concerns. Therefore, we will discuss a V&V process culminating in a lab-test setting, even though some of the data/model can be used for intermediate simulation V&V purposes.

Worldwide, the scenario-based approach is commonly used for the assessment of HAVs. In this approach, the overall driving task during daily trips is decomposed into singular scenarios (e.g., car following, left turn, pedestrian crossing, entering a round-about, etc.) and all tests are based on these identified scenarios. A scenario-based V&V process essentially involves three parts:

1) Deciding which scenarios must be tested for a particular HAV for a particular deployment,
2) Choose and execute the relevant test cases for each scenario,
3) Compute the one or more scores an HAV receives for the tests.

This paper will start from a review of the data collection exercises conducted in EU, China, and the US. We will focus our discussion on a selected scenario (cut-ins) with the intention to show that while there are differences in terms of how

people drive in different regions, they share enough similarities so that the test procedure can be harmonized internationally with the right approach, methodologies, and mindset. The test case parameters may be adjusted to represent actual expected challenges driving on the public roads in different regions. Then, the paper will ask for feedback and inputs on the following:

- Methods and their impact on the requirements towards scenario databases and test cases generation;
- Methods and their impacts on the requirements towards test cases, scenarios, and databases;
- Other approaches, practices and methods is welcome to solidify further insights.

In conclusion, the paper will present a prospect for the near-term future and what contribution is needed in this important assessment.

2. Distinguishing Between Scenarios and Test Cases

A strict distinction must be made between the two terms “scenario” and “test case”. Before addressing both terms, first the concept of “ego vehicle” is introduced. An ego vehicle refers to the perspective from which the world is seen. Here, the ego vehicle refers to the HAV to be tested, that is perceiving the world through its sensors.

According to the European Enable-S3 project [3] [4], a scenario describes any real-world situation that a vehicle out on the road might encounter during its lifetime. A driving trip on the road can be thought to consist of a continuous sequence of scenarios – some of which might overlap. Examples of scenarios include cut-in, car-following, unprotected left turn, facing a pedestrian at a crosswalk, etc. This is independent of the functionalities on board the ego vehicle, or the Operational Design Domain [1] for which the vehicle and its functions has been designed. We consider a scenario to contain a description of [2]:

- Static environment: an environment which does not change through multiple cases of the same scenario. This includes geo-spatially stationary elements, such as the infrastructure layout, the road layout and the type of road. The presence of buildings near the roadside that act as a view-blocking obstruction is considered part of the static environment.

- Dynamic environment: As opposed to the static environment, the dynamic environment changes during the time frame of a scenario. The dynamic environment is described using activities, the way the state of actors evolves over time. In practice, the dynamic environment mainly consists of the moving actors (other than the ego vehicle) or other objects that are relevant to the ego vehicle.

- Conditions: Important for the description of a scenario are also the weather and lighting conditions as these also have an influence on the ego vehicle. For instance, precipitation can have a large influence on sensor performance and vehicle dynamics. Lighting conditions also influence sensor performance.

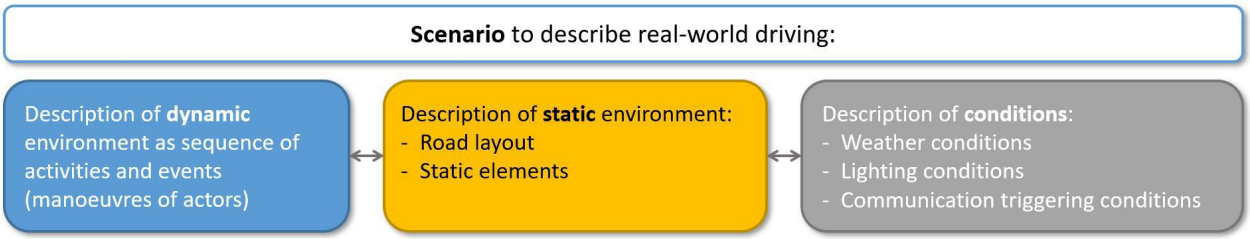

Figure 1 Schematic View on Components that Describe a Real-World Scenario

Dynamic traffic and conditions can be described as a sequence of activities (e.g., lane change, deceleration, light/shadow changes), where multiple activities can take place simultaneously (e.g., the ego vehicle changes lane while during the lane change, a sudden change from shadow to light is perceived).

According to IEEE [5], a “test” is an evaluation of a statement on the system-under-test (the test criteria), under a set of specified conditions (a test case) using quantitative measures (metrics) and a reference of what would be an acceptable outcome (reference). This means that a test case indicates under which conditions each test in an assessment is performed. Figure 2 gives an overview on how test cases are selected:

The orange box in Figure 2 is the set of all possible real-world driving trips. This does not only concern different type of scenarios (e.g., cut-in scenario, or car-following scenario), but also all possible variations in which a scenario may occur (e.g., a vehicle cutting in by aggressively changing lane, or vehicle cutting in smoothly, leaving plenty of space). Scenarios and their variations (test cases) are expected to change for different countries or regions. The objective of

data collection research programs, e.g., StreetWise [6], PEGASUS [7], MOOVE [8], is to compile scenarios and their variations in a database, to capture real-world traffic that richly reflect the local traffic in a particular region. The set of scenarios that are considered known (either because they have been encountered or they are expected to be realistic according to physics – known unknowns) and which are stored in scenario databases are covered by the blue-bounded box area.

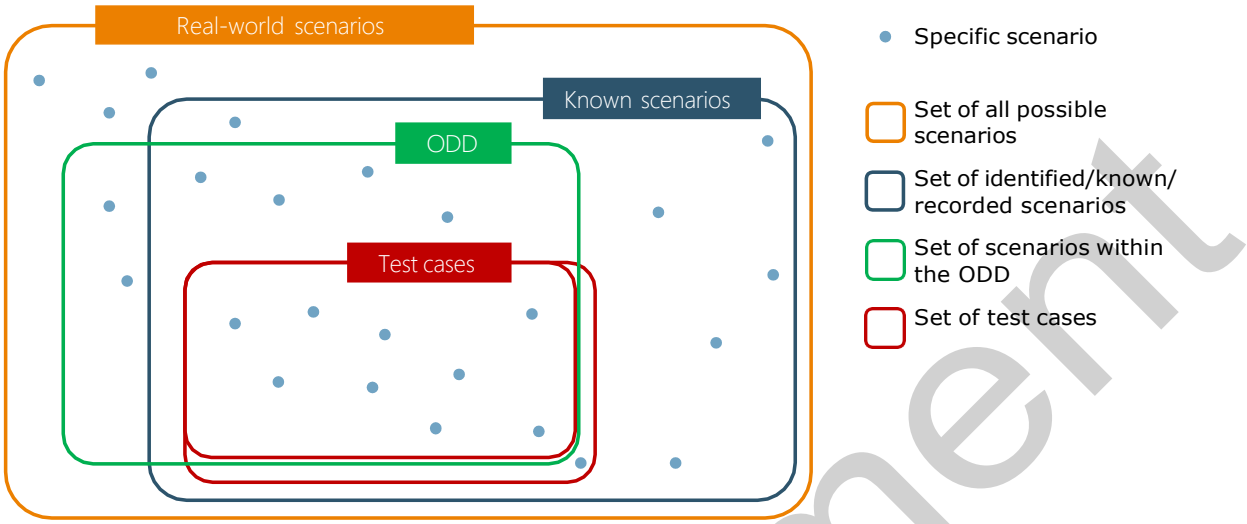

Figure 2 Relation Between Scenarios and Test Cases According to ASAM, the Association for Standardization of Automation and Measurement Systems (www.asam.net)

In general, scenarios provide a view of all different situations that might happen on public roads, but only those that are relevant need to be tested. This is done by making use of a clear description of the ODD (green bounded box) in Figure 2 of the vehicle-under-test following the definition of the ODD according to [4]. The selection of test cases out of the set of known scenarios (blue bounded box) in Figure 2 should be such, that the resulting set of test cases covers the ODD as well. In Figure 2, the set of test cases is indicated by the red bounded box. Here the coverage of test cases is rather poor, as the test cases do not fully cover the known scenarios within the ODD. Ideally, the test cases cover the entire ODD, but some believe that the tests should go beyond the borders of the ODD, hence why we have shown a dashed border line for the set of test cases.

3. Scenario Statistics – Europe Example

The StreetWise scenario database [6] is a scenario database created by partners TNO and AVL. Ten categories of scenarios are captured to represent the most common situations that occur on highways (e.g., lead vehicle decelerating, ego vehicle approaching slower lead vehicle, cut-in scenario in front of ego-vehicle, ego merging in an occupied lane, etc.), but it is by no means complete. The developed methodology is meant to be generic so that it can be used to cover other scenarios (e.g., pedestrian crossing, round-about, etc.) as well.

Each of the scenario categories in the database are described with five to ten parameters so that they can be varied to describe different test cases. The variables are scenario-dependent, and in the case of the cut-in scenario, include ego vehicle speed (i.e. both in driving direction and in lateral direction), lead vehicle speed and acceleration (also in driving direction and in lateral direction where applicable), and the distance between the vehicle that is cutting in and the ego-vehicle at the moment that the other vehicle starts crossing the lane marker to change lane.

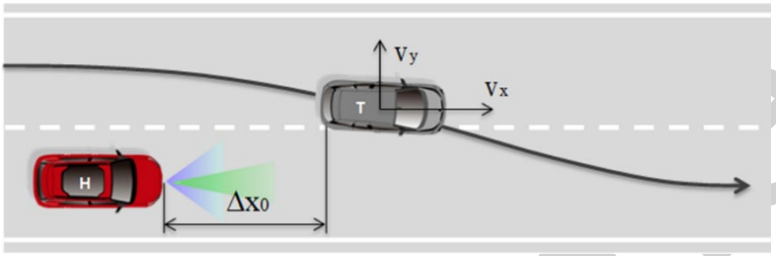

Figure 3 Schematic View of a Vehicle (T) Cutting In on an Ego-vehicle (H)

The following parameter set is currently used to describe a cut-in scenario:

- v_x^{HH}: ego initial longitudinal velocity [m/s]
- Δv_x^{TT}: target initial relative longitudinal velocity with respect to ego [m/s]
- v_y: target average lateral velocity relative to lane over the duration of the lane change [m/s]
- sign vv_{yy}: target lane change direction [-1: from left to right, 1: from right to left]
- THW_{LC}: time headway at start of lane change [s] = $\Delta x x_0 / v_x^{HH}$
- $\Delta x x_0$: distance between target and ego vehicle when target starts crossing the lane marking

These parameters were identified from 6,316 realizations of a cut-in scenario in a data set covering more than 110.000 km of highway driving in Europe. This study not only identified the parameters, but also provided valuable statistical information. This is illustrated in graphs of the probability density functions of the parameters, see Figure 4. The three blue curves in Figure 3 provide the fitted probability density functions for three of the parameters that describe the cut-in scenario.

For the orange graphs, a selection is made to show only a subset of all the possible samples that have an initial time-headway (THW) between 0 and 1 second. One sees from the middle and lower regions of the graph that for that selection, the average lateral speed of the cutting-in vehicle is slightly lower, and the distribution of the relative longitudinal speed of the cut-in vehicle with respect to the ego-vehicle is shifted to the right. This observation agrees with the expectation that for shorter time-headway, the lane change is performed somewhat more carefully (slower), and the gap closing speed between ego and target vehicle is smaller.

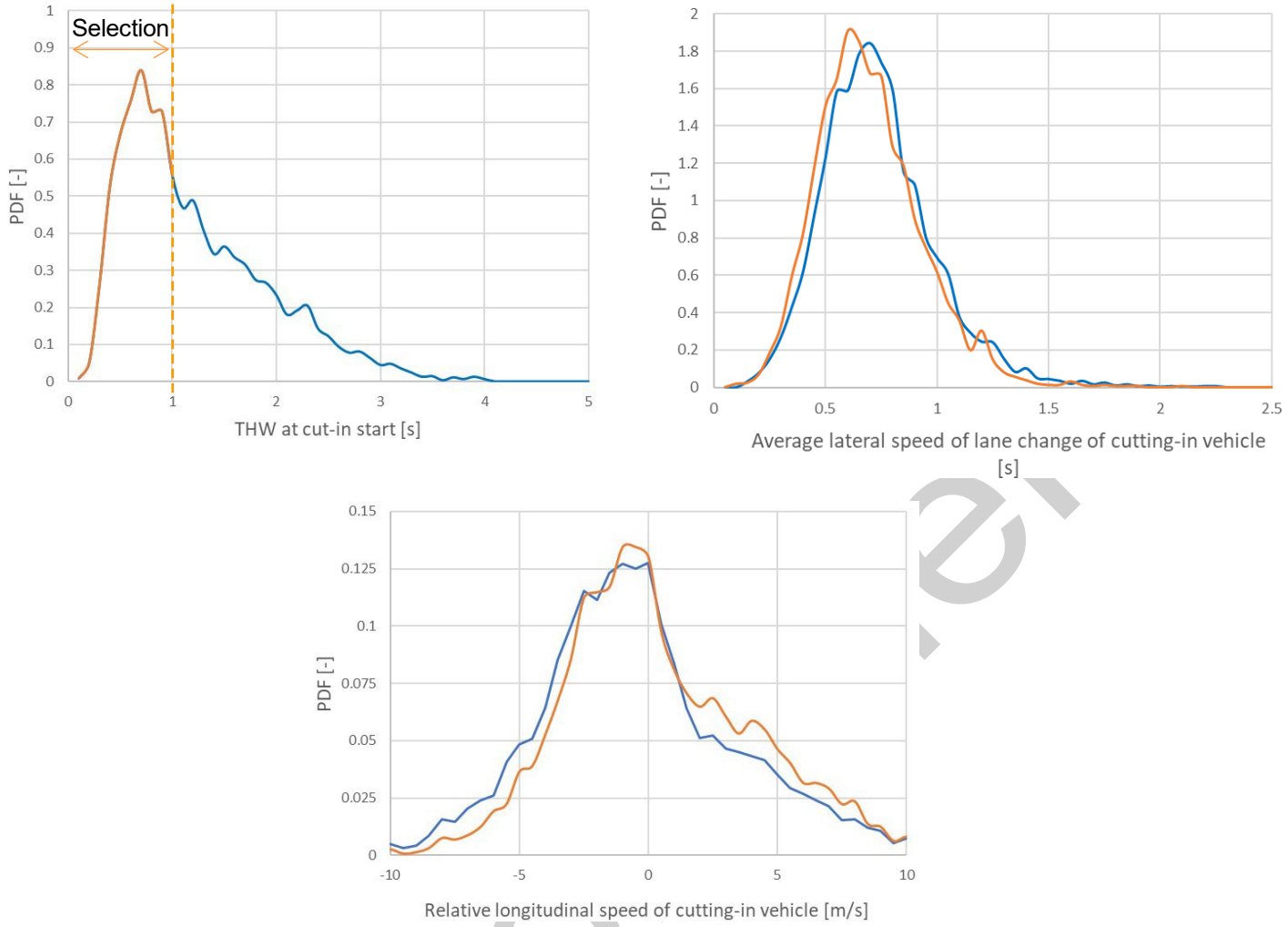

Figure 4 Probability density functions for 3 parameters describing a cut-in (data from EU). The time-headway (THW) between cutting-in vehicle and ego-vehicle at start of the cut-in (upper graph), the average lateral speed of the cutting-in vehicle during the cut

The probability density functions from Figure 4 are exemplary for the parameter distributions that can be drafted when a large number of concrete scenarios (or realizations of scenarios) are identified and characterized. It should be noted that parameters in the description of a scenario are usually correlated. In the above example, this is seen by a change of the distribution of the 'relative longitudinal speed of the cutting-in vehicle' when we only select those cut-in scenarios for which the THW at the start of the cut-in is limited (i.e., to 1.0 second).

The example enables us to provide a definition of 'edge cases' and 'corner cases', which are frequently used terms when discussing test cases for safety assessment. In this paper, we consider a concrete scenario as an edge case, when one of the describing parameters is below the 2.5th percentile or above the 97.5th percentile, and the other parameters are in a range that accounts for 95% of their population. A corner case is considered even more rare than an edge case, with more than one parameter below the 2.5th percentile or above the 97.5th percentile.

4. Scenario Statistics – United States Example

The US cut-in data was extracted from the data of "safety pilot model deployment" project, collected by University of Michigan researchers. About 125 of the test vehicles were instrumented with a Mobileye camera, which captures the motion of a vehicle cutting-in in front of an equipped vehicle. A total of 1.3 million miles of data were collected, among which more than 400,000 cut-in cases were captured.

The statistics of these cut-in cases are shown below. It can be seen from Figure 5 that most cut-in happens with time-headway larger than 2 seconds. The cases with a small time-headway are biased to have positive range rate (i.e., the gap is increasing). During the calculation of the time-to-collision for those cases with negative range rate, it can be seen from Figure 6 that most of the data has time-to-collision larger than 3 seconds, and the duration of the lane-change is longer for highway cut-ins.

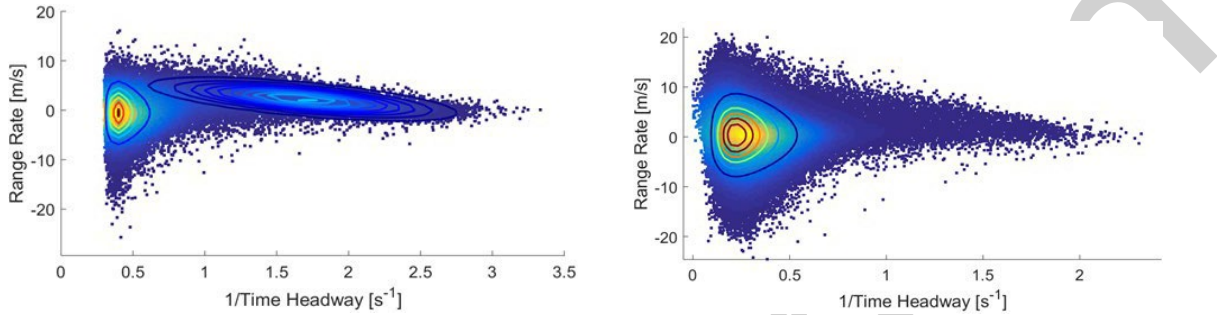

Figure 5 Range Rate vs. Inverse of Time-Headway for US Cut-In Data. (Left: Highway, Right: Local)

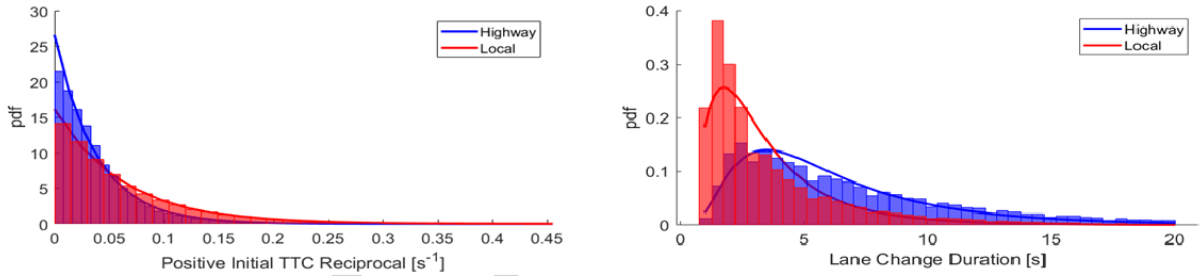

Figure 6 Probability Density Functions of the US Cut-In Data with Negative Range Rate

5. Scenario Statistics – China Example

In 2018, Automotive Data of China Co., Ltd carried out a project to collect more than 100,000 km of naturalistic driving data on China roads. The collected data was processed and extracted a total of 15,645 lane-changing events. [9]

The test vehicles were equipped with an HD video camera, millimeter wave radar, GPS, and an IMU. Among all the 15,645 events, 5,444 of them occurred on highways, 5,734 on city expressways, 4,241 on urban roads, and 226 scenarios occurred on ramps. Based on the collected data, they are classified into 22 defined lane-changing sub-scenarios.

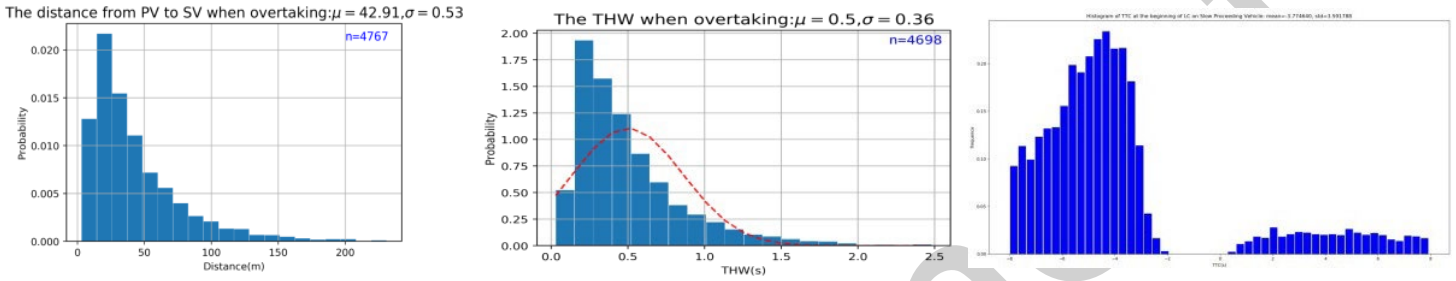

Figure 7 Probability Density Functions of the China Cut-In Data

In addition to collecting data, the research studied why lane-changes under different situations are aborted. The study analyzed the lane-change behavior and proposed four main sub-classes of the scenarios. The distribution and correlation of the parameter under different lane-change sub-classes and different road environments were also analyzed.

6. What We Learned

The three data collection exercises in EU, US and China were all designed for different purposes, with different sensors, driving conditions, geographics, regional specialties and their individual research goals. When we decide to work together to focus on one scenario: lane-change, and in particular, cut-in (i.e., the lane change happens in front of the data-collecting test vehicle at a relatively short distance), we found the data from three regions to be slightly different but share many of the same characteristics. From this comparison, all three regions treated the data with a time-series analysis manner and tries to identify the triggering conditions of a given scenario. While the parameters vary across the regions, the key performance indicator stays the same. Cross-regional differences may come from the difference of road structure, environmental terrain, weather, population density, traffic flow, driving style, on-road vehicle types, etc. More importantly, the probability density functions look qualitatively similar and can be described by the same fitting functions with slightly different model parameters. This is encouraging and hinted that there is high potential to harmonize the test methodology while the detailed execution (test case parameters) can be adjusted according to the region of AV deployment.

GENERATION OF TESTS

Once the naturalistic data is collected, it should be considered how to make it useful for simulation and testing. Test case generation, using a scenario database as input, consists of:

- Describing relevant tests that cover the ODD of the HAV under test, and that trigger the functionalities of the HAV. This can be done by sampling from the parameter distributions for those scenarios and for those parameter ranges that fall within the ODD of the HAV;
- Determining the granularity with which test cases are provided, e.g., what is the grid size in the test case descriptions. The grid size (i.e. distance between test cases) is determined by the number of test cases that is assigned for a specific scenario category and the range of the parameters that needs to be covered according to the ODD. Moreover, some sampling strategies allow for importance sampling, decreasing the 'grid size' in areas of specific interest within the ODD, e.g., in areas for which the risk of collision is higher;

- In addition to test case generation, it is important to determine how to allocate the tests to a test method, so to decide what tests to perform in simulation, what tests to do in Hardware-in-the-Loop (HIL), and what to test on the test track or testbed, etc.

COMBINATION OF DIFFERENT TEST ENVIRONMENTS

Testing an HAV in all the different situations of an ODD requires an intelligent combination of different test environments to develop a robust and safe system within a reasonable time and cost factor.

A three-pillar approach explained in Figure 8 is to simulate all different scenarios of an ODD in a large computing cluster and collect all the data – such as in the case of a cut-in scenario, we might run thousands of experiments by varying some key parameters like the ego vehicle speed versus the time headway at cut-in as shown in the figure below – and from there funnel out the interesting scenarios that are worthy of further inquiry. There will be many "yellow" scenarios (moderately challenging, see Figure 8) – which means they require further analysis, and there will be some "red" scenarios (collision impossible to avoid, see Figure 8) – which means they absolutely need to be looked at more carefully. The exact extraction will depend upon the criteria used to define what is critical, for example it could be as simple as the collision distance dropping below 10 meters beyond a speed of 40mph. This approach further reduces the number of possible scenarios to be tested down to the critical ones. Through this process, the critical scenarios are the ones that you really want to test in more detail.

The "yellow" and "red" scenarios are the perfect candidates to move to the lab testing environment (for example on a chassis dyno or a powertrain dyno lab) – where they can be reproduced and analyzed to allow for the optimization of the automated driving (AD) features until some of them turn "green". As test are also conducted in semi-virtual environments, you will need to accurately simulate what is missing from that environment for the results to be comparable. You can also perform specific test cases which cannot be done in pure simulation due to, e.g., lack of simulation model quality.

At the end there will still be some scenarios here that are "yellow" – meaning they need to further be tested on the proving ground or on the real road. These environments allow for further tuning and optimization of the AD features in most realistic real-world situations.

This described approach allows for smartly balancing the use of the virtual and real-world test platforms that are both expensive and time consuming; and it also enables the use of scaling in simulation and funneling appropriately to the other environment to test the critical scenarios. The time compression from road/proving ground to virtual testing is huge – since you can rapidly reset the simulation and run the next maneuver without delay. It is even higher in simulation due to tests running much faster than real-time and fast change of parameters and weather conditions.

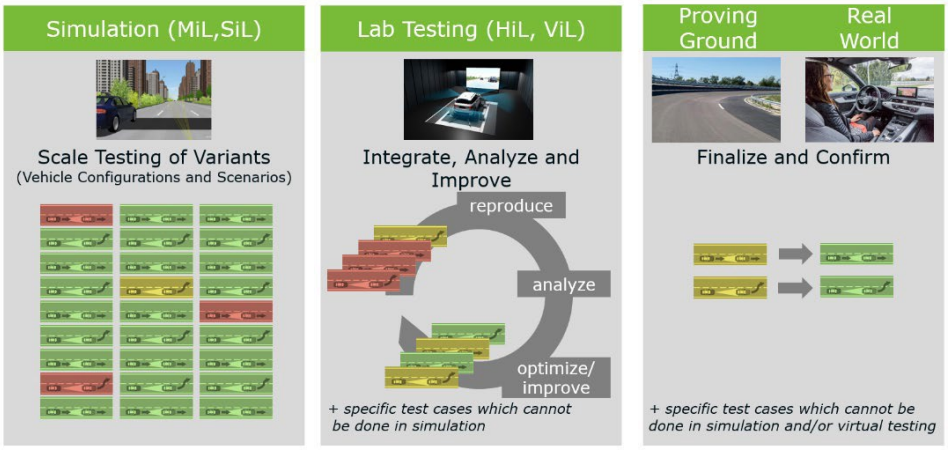

Figure 8 Best Combination of Different Test Environments

TEST CASES FOR SIMULATIONS

The analysis from previous sections showed that high-level automated driving systems testing is a "long tailed question": the most common scenarios are not necessarily the most useful ones. For testing purposes, all three regions are trying to develop methods for scenario identification and characterization on a large scale, e.g., by interpreting the data collected from vehicles performing test drives on public roads. Many approaches are continuously being explored around

the world. In this aspect, a common scenario description language could be very useful. This enables the comparison of scenarios collected by different organizations in different regions of the world.

A second aspect concerns the use of scenarios to provide descriptions of tests as input for simulations ASAM OpenX describes a family of cohesive data formatting standards, which provides a common interface for simulation applications. The ASAM OpenSCENARIO and OpenDRIVE formats support the standardization of the description of a concrete scenario (for replay in a simulation) and of test cases, e.g., resulting from sampling parameter distributions (for massive simulation). The files in such format can be edited, imported, and exported by simulation tools and content editors.

In China, many of the stakeholders are looking for a generic scenario description language that could unify varieties of scenario sources. Figure 9 illustrates the process of using ADScenario tool for generating simulation scenarios.

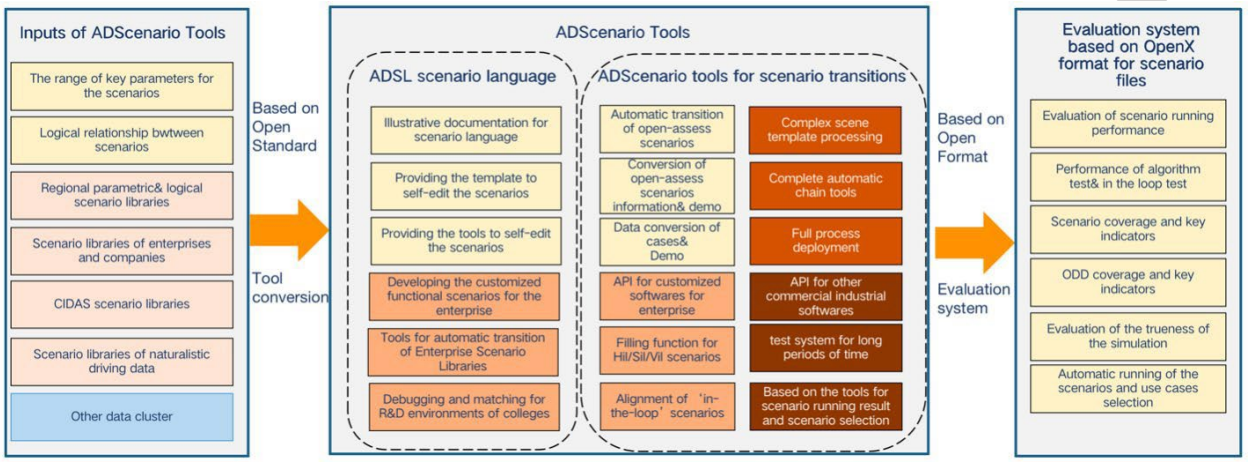

Figure 9 ADScenario is used to generate simulation scenarios at CATARC

With the given framework, a scenario from naturalistic driving data could be transferred into machine readable, simulation ready scenario. However, as mentioned earlier, there are still some missing criteria for using those as test cases. The linkage of those scenarios to a given ODD is then performed before and during the simulation testing for knowledge generation. The following figure shows the procedure of creating test cases from CATARC's scenario database.

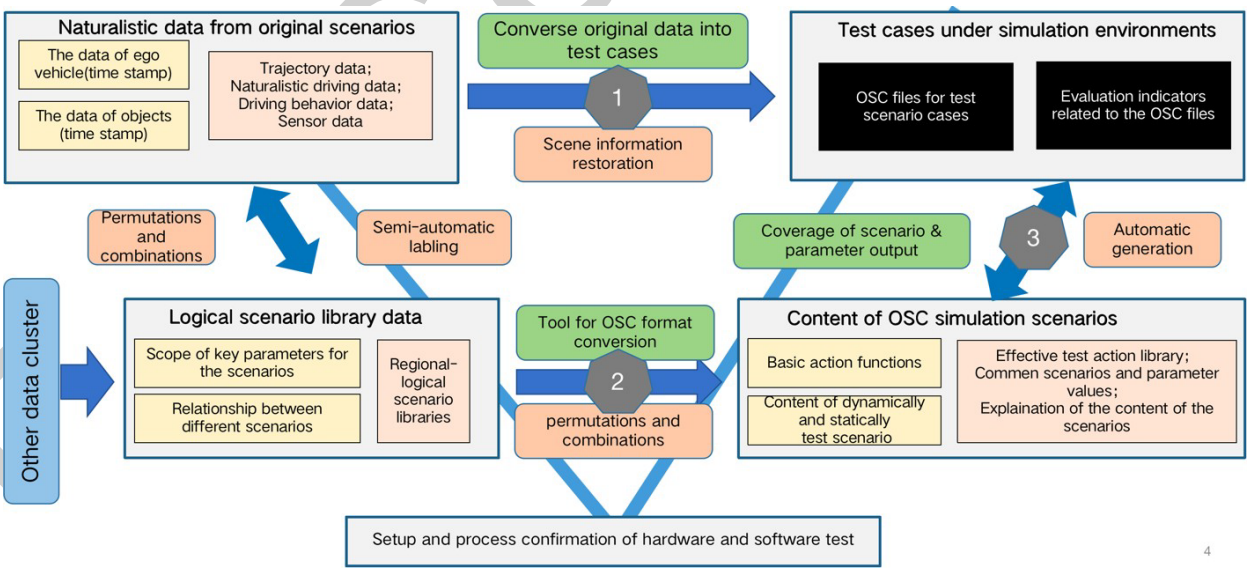

Figure 10 The process going from naturalistic driving data to creating test cases in simulations (CATARC)

The process illustrated in Figure 10 follows more of a "replay" concept used at CATARC, rather than "stochastic sampling" concept used in the US (Mcity). For "replay" we mean that the goal is to recreate a particular driving trip, for which the data is collected. In simulations some of the static/dynamic conditions might be manipulated, such as to simulate adverse weather, sensor failure, etc. This concept is more popular for V&V conducted in simulations. Both of those methods are widely applied and used in scenario-based testing.

In the “replay” concept shown in Figure 10, the naturalistic trip will be queried from the original scenario library to retrieve the trajectory data, naturalistic driving data, driver’s behavior data and sensor data. This data should be chronological and collected from the ego vehicle and objects. In here, the crucial part is labelling the collected data with the accurate scenario labels, sensor labels and behavior labels. This could be beneficial for training reliable machine learning or data mining model for pattern recognition. By using this approach, specific parameter ranges for logical scenario libraries will be generated, including key attributes for the scenarios, logical relationship between scenarios, regional-logical scenario libraries. To make the best practice in utilizing the scenario, key data and key scenario parameters should be tested and implemented within a given HAV system. For each application, the scenario used in the system is the one that covers the given scenario labelling and logical scenario. By applying this scenario permutations, combining logical scenario libraries and other data clusters such as CIDAS scenario libraries, the simulation scenarios described according with domain specific language and a given ontology of the testing setup are required. It covers some of the testing requirements and scenario contents according to Figure 8. In this manner, a related training and testing scenario library will be generated for the validation process. This also enables the possibility of testing with automatic training and end-to-end modelling.

7. Test Cases for Lab Testing

It is generally agreed that HAV developers will use simulations as a tool to make HAV development safer, cheaper, and faster. However, conclusions regarding the safe deployment of HAVs cannot be drawn on the results of simulations only. Physical testing and lab testing are required, both for the validation of the HAV itself and for the validation of the simulation models used for the assessment.

In a lab testing process, assuming that a set of scenarios have been identified for a particular HAV in a particular ODD, the main technical challenge is then to generate the test cases. There are at least four high-level guiding principles for selecting the test case parameters:

1. they should reflect naturalistic road user behaviors;
2. they should be selected stochastically, rather than deterministically;
3. they should be selected in an “accelerated evaluation” fashion; and
4. they should at least cover “the edges and the corners” of the ODD.

To ensure the test case parameters are selected based on actual road-user behaviors, the first step is to collect naturalistic driving data and build a stochastic model, as illustrated earlier in this paper. This is easy to say and hard to do. At the University of Michigan, even with millions of miles of data collected, the university still does not have enough data for all possible scenarios. This is primarily because for some scenarios (e.g., cut-in) the major driving interaction can be captured by on-vehicle sensors, but other scenarios (e.g., round-about) are better collected using road-side sensors looking down at the round-about. Most of the data collected are from instrumented vehicles—a weakness for the University of Michigan projects, as well as for many other data gathering projects, where most of the data are from instrumented vehicles. Figure 11 shows three example scenarios where there was enough data to build useful stochastic models.

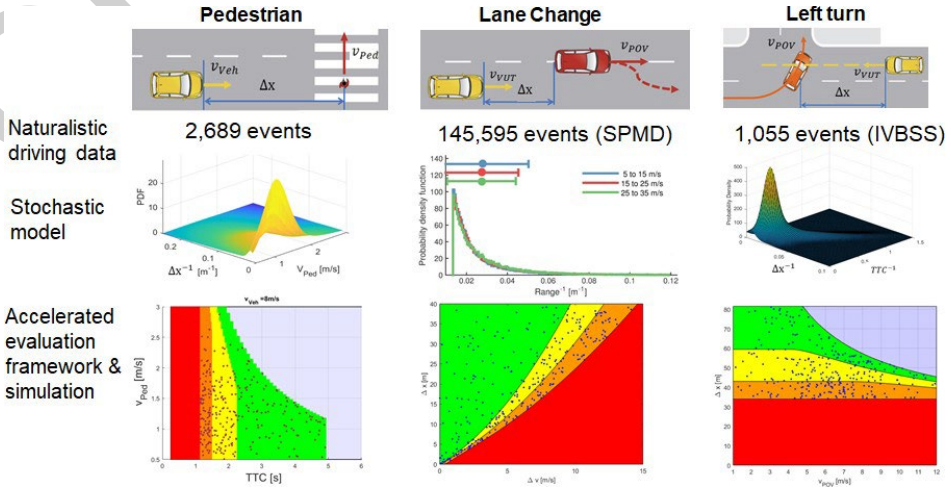

Figure 11 Data and Stochastic Models for Several Scenarios. SPMD is the safety pilot model deployment database, IVBSS is the integrated vehicle bases safety system database.

The statistical models constructed represent what human drivers and pedestrians do when they encounter another vehicle. Depending on the given conditions at the beginning of an event, the space of all possible events is divided into "impossible to avoid" (red), "possible to avoid," and "trivial" (blue) regions first, and the "possible to avoid" region is further divided into three sub-regions: orange (highly challenging, or hard), yellow (moderately challenging, or medium) and green (low challenging, or easy). Take the pedestrian crossing scenario as an example. "Impossible to avoid" represents the cases when the pedestrian suddenly dashes in front of the vehicle, and there is simply no time for the vehicle to brake or swerve to avoid a crash. "Trivial" captures the cases when the pedestrian walks in front of the vehicle at such a far distance that the vehicle does not need to take any action. The AV can drive at its current speed, a crash or near-miss (defined by a minimum separation distance or time-margin) will not occur. All cases in between "impossible" and "trivial" is "possible to avoid", which then needs to be further divided. The technique behind separating the "possible" space into different regions is control reachability analysis, and assumed delay/maximum braking capabilities of the AV. The details can be found in reference [16].

In Figure 11, the lowest row of the table the dots (each representing a stochastically sampled test case) are evenly distributed in the easy, medium, and hard regions, and some are approaching the red-zone (the impossible region). In combination with Figure 12, it becomes clear how one achieves "accelerated evaluation". For accelerated evaluation, the lab testing explores riskier test cases much more efficiently than what would be encountered driving on public roads naturalistically. If the statistical model is sampled naturalistically, i.e., based on observed data collected from the public roads, then all the 300 samples fall within the "easy" region (see the left plot of Figure 12). This obviously would be statistically close to what one would experience on the public roads. However, this is not efficient for a regulation/approval test, i.e., all tests are easy and uneventful. In order to achieve "accelerated evaluation", as an example, sampling 1/3 from the "easy" region, 1/3 from the "moderate" region and 1/3 from the "hard" region. The same 300 samples would expose the AV to much more difficult test cases, achieved within tens of total miles driven, instead of millions of miles in order to catch those corner cases (close to the boundary of orange and red regions). The theoretical basis of the accelerated evaluation concept is rigorously presented in [17] and subsequent papers, and what is presented above is a simpler explanation and a process that is easy to execute in practice.

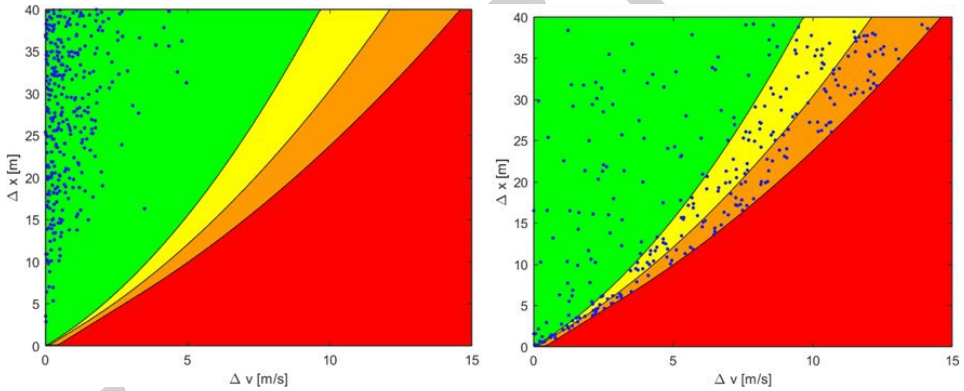

Figure 12 Vehicle Cut-In Cases. Sampled naturalistically (left) vs. sampled evenly in the three (possible) colored regions (right). Each dot represents a random sample.

In order for the general public to accept HAVs, safety must be demonstrated, and we must have a clear process that is open, rigorous, repeatable, and ever-improving. After the test cases are selected, they must be executed accurately and reliably. By dividing the "possible and avoidable" regions into three sub-regions based on levels of defined challenge, it is then possible to select test cases that are different, but fair. When the exact test conditions are known in advance, such as the approach used today through "test matrices", there is a higher chance companies will focus solely on passing the test and not on improving the true performance. Many government agencies are now aware of this practice and have vowed to prevent it in future government-sanctioned tests. Our proposed method achieves that goal.

8. Robust Validation and Release Decisions

After each test case in simulations or lab testing, a "score" must be calculated, and compiling the results from all test cases under all scenarios, a decision must be made about whether the HAV passed the test, and whether it can be deployed or further tested on the public roads. While most of the discussions converge to vague ideas like "HAVs must be safer than average alert human drivers", those metrics are hard to measure and certify, and need billions of miles to certify with high confidence, which is feasible only through simulations but not lab testing. Scenario-based lab tests today can merely check for "behavior competence" and the test cases must be selected to approach the performance boundaries that we expect the HAVs to experience in real-world situations. We envision that the HAV developers will use simulations to cover millions/billions of miles, and the lab tests overseen by government agencies can only check and verify behavior confidence.

For both simulations and lab tests, while safety is and should remain to be the top priority, being safe alone is not sufficient for a HAV. Take an unprotected left-turn and entering a round-about as an example, a HAV that is safe but fails to take many safe gaps is unacceptable. Similarly, a vehicle that brakes unexpectedly and too harshly than "typical" human driven vehicles, especially when there is no clear reason to do so, can be a nuisance or hazard to other road users or the onboard passengers. In [18], the term "roadmanship" was defined as "the ability to drive on the road safely without creating hazards and responding well (regardless of legality) to the hazards created by others." After each V&V test scenario, in addition to assessing safety, the "roadmanship" will also be evaluated. The HAV will be penalized for driving too slowly, or exceeding the speed limit, weaving too much, or braking too hard for no obvious reasons.

9. Conclusion and Prospect for the Future

Verification and validation of HAV is a pressing issue and must be addressed by governments in EU, Asia and the US in the near future. The paper reflects a wealth of insights from expert practitioners towards the use of scenario databases and the generation of test cases for the validation and verification of HAV safety. With different objectives for different stakeholders, the applied methods with a combination of virtual simulation, physical testing, and road testing, show a wide variety, and so does the value derived from these methods.

The examples illustrate that the scenario database and the resulting test cases derived from the scenarios need to fulfill a multitude of requirements. These requirements can be articulated in the following manner:

- How to construct them?
- How to exploit them using different methods?
- How to interpret results using them?

There is not one harmonized common database of test cases that can fulfill all requirements; however, methods are available that enable the exploitation of scenarios and the generation of test cases to serve the different objectives across stakeholders and across regions with a sensible degree of harmonization, federation, and aggregation.

It is encouraging to see that while there are large differences in data distribution and characterization across regions due to population density, road networks, traffic density, traffic rules and resulting operational domains, these differences can be quantified and used for safety assessment according to the locally applicable regulatory constraints.

The paper makes clear that, eventually, we will need federated scenario statistics and scenario reference sets that can serve virtual simulation, testing on test tracks or lab testing, and public road tests on a global scale. This is an ambitious goal for which we will need:

- Your input to the insights and findings across the examples in this paper;
- Feedback to the methods and suggestions for other relevant practices;
- Contributions to this objective via the IAMTS membership.

10. Contact Information

To learn more about the International Alliance for Mobility Testing and Standardization™, please visit http://iamts.org

Contact: info@iamts.org

11. Contributors

Huei Peng, Mcity

Zhixin Wu, CATARC

Bolin Zhou, CATARC

Olaf op den Camp, TNO

Thomas Guntschnig, AVL

Charlie Cheng, SAE ITC

中国汽车技术研究中心有限公司
China Automotive Technology and Research Center Co., Ltd.

CATARC is the leading third party, simulation validation organization in China. It has the biggest scenario database in China for validation and verification purposes and provides related toolchain and datasets for ADS testing.

TNO innovation for life

TNO Automotive is widely recognized by industry and governments all over the world as a valuable knowledge partner with unique expertise, methodologies, and facilities for the development of innovative solutions that make our vehicles more safe, efficient and sustainable.

AVL

AVL is the world's largest independent company for the development, simulation and testing of powertrain systems, their integration into the vehicle as well as new fields like ADAS/AD and Data Intelligence.

M city
UNIVERSITY OF MICHIGAN

Mcity is a public private partnership that brings together industry, government, and academia to advance transportation safety, sustainability, and accessibility for the benefit of society.

SAE ITC
An SAE International Affiliate

SAE ITC enables public, private, academic and government organizations to connect and collaborate in neutral, pre-competitive forums.

12. References

[1] SAE On-Road Automated Driving (ORAD) committee, “SAE J3016, Taxonomy and Definitions for Terms Related too On-Road Motor Vehicle Automated Driving Systems,” SAE International, 2018.

[2] E. de Gelder, O. Op den Camp and N. de Boer, “Scenario categories for the assessment of automated vehicles,” 2020. [Online]. Available: http://www.cetran.sg/publications.

[3] S. Kalisvaart, Z. Slavik and O. Op den Camp, “Using Scenarios in Safety Validation of Automated Systems.,” in Validation and Verification of Automated Systems., Springer, 2020.

[4] A. Leitner and M. Paulweber, “Testing & Validation of Highly Automated Systems,” in ENABLE-S3 Summary of Results, 2019.

[5] E. Stellet, M. R. Zofka, J. Schumacher, T. Schamm, F. Niewels and J. M. Zöllner, “Testing of Advanced Driver Assistance Towards Automated Driving: A Survey and Taxonomy on Existing Approaches and Open Questions,” in IEEE 18th International Conference on Intelligent Transportation Systems, 2015.

[6] H. Elrofai, J.-P. Paardekooper, E. d. Gelder, S. Kalisvaart and O. Op den Camp, “StreetWise position paper, scenario-based safety validation of connected and automated driving,” 2018. [Online]. Available: www.tno.nl/STREETWISE.

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

[8] G. Thiolon and A. Bracquemond, “Real world driving scenario identification for AV functional safety,” in Autonomous Vehicle Test & Development Symposium, Stuttgart, 2018.

[9] Systems, Association for Standardization of Automation and Measurement, „ASAM Open Simulation Interface,” [Online]. Available: https://www.asam.net/standards/detail/osi/.

[10] UN Economic Commission for Europe, Inland Transport Committee, World Forum for Harmonization of Vehicle Regulations, „UN Regulation on uniform provisions concerning the approval of vehicles with regards to Automated Lane Keeping System,” in ECE/TRANS/WP.29/2020/81, Geneva, 2020.

[11] E. de Gelder, A. Khabbaz Saberi and H. Elrofai, “A Method for Scenario Risk Quantification for Automated Driving Systems,” in 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV), Eindhoven, 2019.

[12] J.-P. Paardekooper, S. van Montfort, J. Manders, J. Goos, E. de Gelder, O. Op den Camp, A. Bracquemond and G. Thiolon, “Automatic Identification of Critical Scenarios in a Public Dataset of 6000 km of Public-Road Driving,” in Conference on the Enhanced Safety of Vehicles, Eindhoven, 2019.

[13] E. de Gelder, J.-P. Paardekooper, O. Op den Camp and B. De Schutter, “Safety assessment of automated vehicles: how to determine whether we have collected enough field data?,” Traffic Injury Prevention, vol. 20, pp. S162-S170, 2019.

[14] P. Wimmer, M. Düring, H. Chajmowicz, F. Granum, J. King, H. Kolk, O. Op den Camp, P. Scognamiglio en M. Wagner, „Toward harmonizing prospective effectiveness assessment for road safety: Comparing tools in standard test case simulations,” Traffic Injury Prevention, vol. 20:sup1, pp. S139-S145, 2019.

[15] J. G. Taiber, S. Hinze, C. Rösener, J. Tintinalli, T. Bock, S. Rößner, I. A. Zlocki, “DIN SAE SPEC 91381, Terms and Definitions Related to Testing of Automated Vehicle Technologies,” DIN SAE SPEC, 2019.

[16] Wang, Xinpeng, Huei Peng, and Ding Zhao. "Combining Reachability Analysis and Importance Sampling for Accelerated Evaluation of Highway Automated Vehicles at Pedestrian Crossing." ASME Letters in Dynamic Systems and Control 1, no. 1 (2021).

[17] Zhao, Ding, Henry Lam, Huei Peng, Shan Bao, David J. LeBlanc, Kazutoshi Nobukawa, and Christopher S. Pan. "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques." IEEE transactions on intelligent transportation systems 18, no. 3 (2016): 595-607.

[18] J.Holzinger, A.Leitner, M.Nager, H.Schneider, “Scenario-Based Validation Framework for ADAS / AD for multiple execution environments - MIL / SIL / HIL / VIL”, 26th ITS World Congress Singapore (2019)