Conducting the Mcity ABC Test:
A Testing Method for Highly Automated Vehicles

HUEI PENG, PhD
Director, Mcity
Roger L. McCarthy Professor of Mechanical Engineering

Contents

INTRODUCTION

The past few years have seen a Cambrian-like explosion of companies, universities, and other research centers around the world developing highly automated vehicles (HAVs), or vehicles that can operate without a human driver. For purposes of this paper, HAV refers to vehicles that operate with Level 4 or Level 5 automation features, as defined by SAE International, a global association of engineers and technical experts from the aerospace and automotive industries. A simpler way to understand HAVs is that they are vehicles chauffeured by a robotic driver; you, the human, are a passenger.

While HAVs have the potential to improve transportation safety, convenience, and accessibility, a key challenge today is a lack of trust by many people in the actual safety of HAVs. Currently, most HAVs are safety tested through a combination of software simulation, closed-track testing, and on-road testing. The National Transportation Safety Board (NTSB) recently published the results of its investigation into a 2018 fatal crash involving an automated Uber vehicle in Arizona. [1][2] The report called for a tighter safety process by the National Highway Traffic Safety Administration (NHTSA), the US Department of Transportation unit that oversees motor vehicles. Today, any HAV developer can apply and, if approved, test on public roads in many states with little

oversight. While NHTSA asks for a voluntary safety report [3], only a small number of companies have submitted one. In addition, there is no independent safety verification required before beginning public road testing, which is a crucial element for improving the public's trust in automated vehicle technology.

MCITY ABC TEST

In a white paper published in January 2019 [4], Mcity introduced the concept of the "Mcity ABC Test," a three-pronged approach for testing HAVs inside a closed test track, before public road testing and deployment. The three components are **A**ccelerated evaluation, **B**ehavior competence, and **C**orner cases, each of which achieves a different evaluation outcome. Accelerated evaluation focuses on the most common risky driving situations, behavior competence on demonstrating the ability to be safe in a wide array of scenarios, and corner cases on pushing the limit toward the boundary (corners) of the operational design domain (ODD) of the HAV. The concept of ODD is likely to be with us for a long time—it defines the conditions under which an HAV will operate. When the HAV operates at "SAE Level 5," the ODD is whenever, whatever and wherever human drivers can safely drive. Before an HAV achieves Level 5, its operation is somehow constrained, e.g., speed limited, weather limited, geo-fenced, etc.

With this white paper, we go one step further to discuss the main challenges of conducting these tests, and the steps for completing them. We also present a concept for how the test results could be used to score the performance of the HAV using some form of an ordinal ranking system. We focus our discussion on the behavior competence tests, but the majority of the content applies to accelerated evaluation and corner case tests as well.

BEHAVIOR COMPETENCE TEST METHODOLOGY

Here is an outline of the steps necessary for conducting a behavior competence test.

1. Behavior Competence Test Scenarios

The intent of the behavior competence test is to verify that the HAV is capable of handling a wide array of driving scenarios safely. Several organizations have published driving scenarios that they think should be tested. After reviewing the literature, Mcity

researchers compiled a list of 50 scenarios [4]. The exact number of scenarios is not our focus since the number is likely to increase as testing procedures mature. In addition, the subset of scenarios that should be included to test a particular HAV needs to be defined based on the operating domain of the vehicle. As an example, for GPS-following, fixed-route (though still Level 4) shuttles, an unprotected left turn, a steep uphill climb, or a roundabout may not be part of the deployment route and therefore may not need to be tested. Once the subset of the behavior competence scenarios is defined, two questions remain:

- How to choose the parameters of the test cases? There should be multiple test cases for each scenario to ensure adequate coverage.
- How to score the HAV based on the results?

2. Guiding Principles for HAV Behavior Competence Tests

Before reviewing the technicalities in selecting the parameters of the test cases, it is important to first outline the principles of how these cases should be generated and how the results should be used to compute a score for the HAV being tested. These principles include:

(i) The test scenarios and test case parameters should be based on the ODD of the HAV. Vehicles that use Level 4 automation features are meant to operate within certain constraints. The scenarios and test case parameters therefore must be selected only within the bounds of the ODD, and not beyond. In other words, all test cases selected for the HAV need to be foreseeable and preventable within the intended deployment bound. Unpreventable test cases are not useful in terms of assessing the behavior competence of HAVs.

(ii) In choosing the parameters of the test cases for each scenario, the behavior of other human road users (drivers, pedestrians, cyclists) should be sampled stochastically instead of deterministically. Stochastically, or randomly, varying test cases encourages HAVs to have real behavior competence rather than just focusing on successfully completing the pre-selected "test matrices".

(iii) The test case parameters ideally should be selected based on naturalistic road user data. There are significant differences between road user behavior in different countries, or even different regions within a country. Therefore, naturalistic behavior statistics can be customized based on the HAV deployment target area, provided such localized driver behavior data is available. Selecting test case parameters in this way ensures that the tests are statistically similar to what the HAV will experience in the real deployment.

(iv) The test cases need to be associated with a clearly defined challenge level (e.g., high, moderate, low). All HAVs being tested from multiple manufacturers should be given an equal number of test cases in each of the challenge levels so that they are equally difficult, meaning the test cases for HAVs from different manufacturers can be different but the test must be fair. In this white paper, we separate the test cases into three challenge levels. Finer challenge levels can be used following the concept illustrated in this paper.

(v) Test cases must be executed precisely. For many testing scenarios, the important test case parameters are relative distance and relative speed at, or in a short period before, the "conflict zone." Regardless of the exact behavior of the HAV being tested, the challenge vehicle (or the pedestrian proxy) needs to execute the test case precisely.

(vi) The test results should be used to determine the performance of the HAV in terms of both safety and "roadmanship." The concept of roadmanship is relatively new and refers to driving behavior that is statistically "normal," or similar to most human drivers. In a report about automated vehicle safety published in 2018 by the Rand Corporation [5], the term roadmanship is 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." Consider unprotected left turns or entering a roundabout as examples. An HAV that is safe but fails to take advantage of safe opportunities to turn or merge is unacceptable because it impedes traffic flow and may induce unsafe behaviors from other human drivers. Similarly, a vehicle that brakes unexpectedly and more harshly than typical human-driven vehicles, especially when there are no justifiable or obvious reasons to do so, can be a nuisance or even a hazard to other road users or onboard passengers. Without proven roadmanship, self-driving cars may never earn consumer trust and confidence at a level necessary to support their widespread adoption. Nor will they fulfill their promise to improve traffic safety, and enhance lives by conserving energy and increasing accessibility to transportation.

3. The Behavior Competence Test Procedure

Following the guiding principles outlined in the previous section, we propose the following procedure to execute behavior competence testing.

(i) Depending on the ODD, select a subset of the library of (50) scenarios that are relevant for the HAV to be tested.

(ii) For the selected subset of scenarios, construct a model that represents naturalistic behavior of other human road users (drivers, pedestrians, cyclists). If data from a statistically significant number of naturalistic driving trips are available, this step may be based on the empirical data. Otherwise, assumptions can be made based on other scenarios, data or models, and semi-empirical models can be built and used.

(iii) For each scenario, use reachability analysis techniques to separate all test cases into "impossible," "hard," "moderate," "easy," and "trivial" cases. Reachability analysis is the process of computing what the reachable states are for a given scenario. Details of how this is done are shown in the next section.

(iv) Sample from the models and execute the test cases.

(v) Record the test results and score the HAV.

4. Roadmanship and Why Safety Alone is Not Enough

How do we program roadmanship into HAVs, and how do we measure it as a performance metric complementary to safety? While safety is and should remain the top priority, safety alone is not sufficient. Roadmanship can be measured by how "normal" the behavior of the HAV is compared with the human driver population. The more an HAV behaves in a "normal" – and thus predictable – fashion, the safer other human drivers will be when interacting with it. If a large quantity of naturalistic driving data is available, a statistical model can be constructed and the roadmanship can be measured by checking where the HAV's behavior falls within the statistical model. While we advocate for measuring both safety and roadmanship, we do not believe they should be treated equally. Safety cannot be compromised, and therefore may be best measured on a pass/fail basis, assuming all test cases were reasonable and preventable. Roadmanship, on the other hand, is an attribute that can allow some variation, and may be better measured on a relative scale, perhaps using a 5-star rating system.

While the miles driven by prototype HAVs are still too low (in the tens of millions of miles range altogether) to draw concrete conclusions, early statistics show that HAVs have been involved in crashes more often than human-driven vehicles. Most of these crashes involved an HAV being struck from behind or side-swiped, and in many cases occurred when the HAV was fully stopped. There are three explanations for the higher frequency of crashes compared with human drivers: (i) under-reporting of minor crashes involving human-driven vehicles, (ii) the HAVs were not behaving "normally" in the eyes, and minds, of other human drivers, or (iii) there was an unsafe change in behavior by human

drivers (a.k.a. "bullying") around the HAVs. Among the three, only (ii) is under the full control of the HAV designer and therefore will be what we focus on.

5. Examples Using Three Behavior Competence Scenarios

In this section, we use three behavior competence scenarios as examples to illustrate the process of analyzing, designing, and executing the tests for each scenario, and the process of using the test results to score the performance of the HAV. The three selected scenarios are cut-ins, unprotected left turns, and pedestrian crossings (Figure 1). In Figure 1, the black car is the HAV under test.

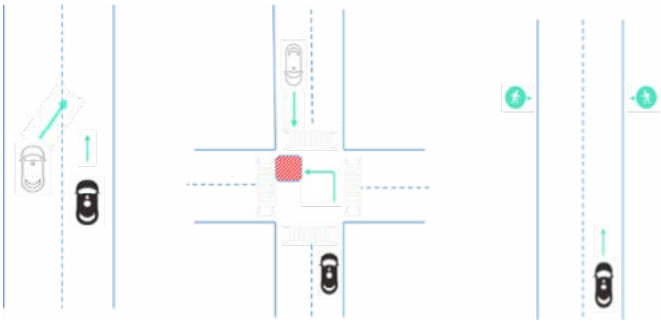

Figure 1: The three selected scenarios for in-depth discussion

For each identified scenario, an accurate naturalistic driving model can be constructed only if there is data from thousands of events. This is because when there are two road users (the HAV and the challenging human driver/pedestrian/cyclist), there are usually two or more independent variables that describe the specific case. Figure 2 shows the number of cases, source of data, and the representative stochastic model we constructed for each of the three scenarios. The best way to collect training data for each scenario varies by scenario. For example, cut-in events are better collected by a vehicle equipped with on-board sensors such as a forward-facing camera, radar or lidar. Unprotected left-turns and pedestrian-crossing data, however, are better collected from a camera placed above the roadway looking down at an intersection, or by using a drone. In Figure 2, SPMD refers to the Safety Pilot Model Deployment database [6], and IVBSS refers to the Integrated Vehicle-Based Safety Systems database [7], both collected and managed by the University of Michigan Transportation Research Institute. Also in Figure 2, VUT stands for vehicle under test, such as from an HAV manufacturer. POV indicates the "Primary Other Vehicle," also known as the challenging vehicle, or the vehicle under the control of the test conductor. The pedestrian crossing data is collected from an open source camera, and we use a deep neural network-based machine vision algorithm to detect the position and velocity of the vehicle and the pedestrian (Figure 3). The objects are then tracked

and their motions smoothed through signal processing techniques. Obviously, our data collection sources and processes are not unique. Newer technologies, such as drones, may be a better way of collecting data more accurately and quickly.

The statistical model constructed represents what human drivers and pedestrians do when they encounter another vehicle. Depending on the given condition at the beginning of an event, the space of all possible events is divided into "impossible" (red), "possible," and "trivial" (blue) regions first, and the "possible" region is further divided into three sub-regions: orange (highly challenging, or hard), yellow (moderately challenging) and green (low challenge, or easy). Take the pedestrian crossing scenario as an example. "Impossible" represents the cases when the pedestrian suddenly dashes in front of the vehicle, and there is simply no time or physical possibility for the vehicle to stop or swerve to avoid a crash. "Trivial" captures the cases where the pedestrian walks in front of the vehicle at such a far distance that the vehicle does not need to take any action. As long as it drives at its current speed, a crash or near-miss (defined by a minimum separation distance or time-margin) will not occur. The above discussion applies equally well to the other two example scenarios. Note that at first glance the cut-in, or lane change, scenario does not seem to have any trivial (blue) region. In fact, all the space to the left of the green region is the trivial region, meaning if the cut-in vehicle drives faster than the HAV under testing, the HAV does not need to take any action, making it a trivial challenge.

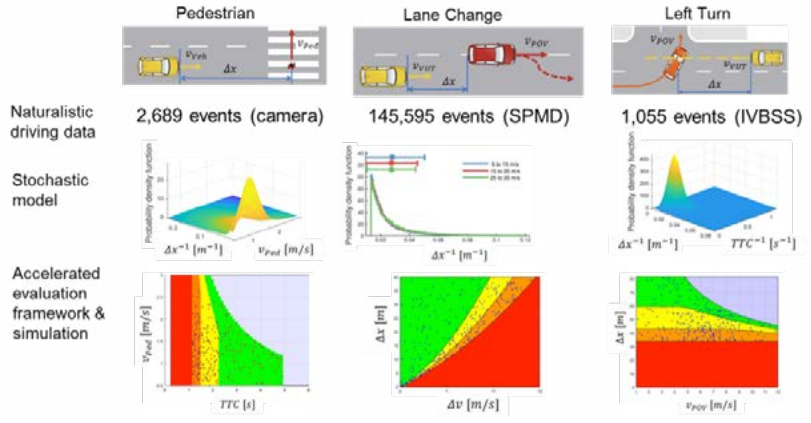

Figure 2: Data collection, model fitting and test case generations

Figure 3: Object detection, tracking and pedestrian-crossing conflict modeling

The next major task is to divide the "possible" region into the three sub-regions: Here we use orange, yellow, and green to represent hard, moderate and easy test cases. The lines dividing these regions are selected based on our observations from how human drivers operate their vehicles. In a 2005 paper [8], data from 107 drivers traveling over 110,000 miles were analyzed. It was found that when human-driven vehicles slow down, 99 percent of the time the deceleration level is below 0.23g, 99.9 percent of the time it is below 0.41g, and 99.99 percent of the time it is below 0.65g. A test vehicle on a dry surface can easily achieve a higher level of deceleration, but that is not a level that most drivers feel is comfortable. Most human drivers anticipate other road users' actions and therefore do not brake heavily often. The following deceleration levels and assumed detection/reaction time were used to construct the lines that separate the regions as follows:

- The line between the red and the orange regions assumes a 0.2 second perception/reaction time delay and a 0.65g deceleration
- The line between the orange and the yellow regions assumes a 0.4 second perception/reaction time delay and a 0.41g deceleration
- The line between the yellow and the green regions assumes a 0.6 second perception/reaction time delay and a 0.23g deceleration

In the above, the perception/reaction time we expect from the HAV is much shorter than reported human reaction time. Human reaction time is highly variable, but was found [9] [10] to be as short as 0.7 seconds, with the average at around 1.3 seconds, and as long as a few seconds. However, safe human drivers' anticipatory behavior to a crossing pedestrian, or to a cut-in vehicle, as examples, compensates for their longer delays. The much shorter reaction time (0.2, 0.4 and 0.6 seconds) used in the above reachability analysis may need to be adjusted based on the nature of the scenarios and should be treated as suggested ball-park numbers.

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. Each

HAV under test is given the same number of easy, moderate, and hard test cases, but the exact test case parameters are not revealed in advance. This ensures that the HAV is prepared for the given scenario over the whole region of "possible and avoidable," instead of only preparing for the "test matrix," a common practice for testing Level 1 and Level 2 automation. When the exact test conditions are known in advance, 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.

Behavior Competence Scenarios	Test Case Parameters
Perform low-speed merge	Time margin, speed margin
Adjust position for vehicles encroaching in lane	Longitudinal offset, lane encroachment
Detect and respond to encroaching oncoming vehicles	Speed margin, lane encroachment
Perform car following (including stop and go)	Speed margin, deceleration level
Detect and respond to stopped vehicles and stationary obstacles	Longitudinal offset, road curve
Detect and respond to lane changes (cut-ins)	Time margin, speed margin
Navigate intersections and perform left and right turns	Time margin, speed margin
Navigate roundabouts	Time margin, leg of entrance
Navigate a parking lot respond to reversing vahicles and locate spaces	Time margin, occlusion
Detect and respond to non-collision safety situations (e.g. vehicle doors ajar)	Time margin
Respond to vehicles breaking rules at traffic lights	Time margin, speed margin
Navigate environments with occluded view	Time margin, occlusion
Detect and respond to golf carts	Speed margin
Make appropriate right-of-way decisions at crosswalks (pedestrians + bicycle)	Ped/bike speed, time margin
Detect and respond to pedestrians in road (not at intersection or crosswalk)	Ped speed, time margin
Keep safe distance from pedestrians and bicyclists on side of the road	Ped/bike speed, lateral offset

Figure 4: Key test parameters of the example behavior competence tests

After the test cases are selected, they must be executed accurately and reliably. Figure 4 shows the key test case parameters that must be accurate. It can be seen that "time margin" and "speed margin" appear in many scenarios. This is because the ability to control the challenging vehicle to arrive at the conflict point at the right time with the correct relative speed is important. We use three HAV-capable Mcity test vehicles (Figure 5) for our tests to serve both the role of the HAV and the challenge vehicle. For pedestrian and cyclist testing, we use a low-profile platform carrying human proxies to achieve precise motion controls. In order to control the challenge vehicle accurately, we install real-time kinematic (RTK) kits on both the HAV being testing and the challenge vehicle. With RTK, the position and velocity of both vehicles can be measured accurately. The position and velocity information are then exchanged through dedicated short-range communication (DSRC), similar in some ways to WiFi. The RTK-DSRC plus a well-designed closed loop control system, or servo loop, are the key components for accurate control of the tests.

Figure 5: Executing the test cases precisely (a cut-in example is shown on the left) using the Mcity test vehicles

SCORING THE TEST RESULTS

To accurately score the HAV's performance, recall that the first step was to determine the subset of scenarios that an HAV should be tested under, based on the target deployment environment and the intended operating domain. Assuming further that each scenario will be tested *m* times in an enclosed test facility such as the Mcity Test Facility, one can then determine the split of easy, moderate and hard test cases (e.g., [m/3, m/3, m/3], or [m/2, m/3, m/6], etc.) Note that the split has implications in "accelerated evaluations" [11], if the intent is to compute risk exposure quickly. Other than that fact, the selection of the split ratios is somewhat up to the tester, as long as all HAVs are tested under the same fixed split ratio so that fair comparison among them is possible. Recalling further that we will allow no compromise with regards to safety, the HAV's safety performance should be graded based on a pass/fail basis. If the minimum distance or time separation is shorter than an agreed upon threshold, then the HAV is judged to have failed the test.

Grading for Roadmanship

Currently, there is no clear consensus on how roadmanship should be realized, or exactly how many test scenarios should consider both safety and roadmanship. Nevertheless, several concepts are agreed upon by most of those with whom we engaged on this topic:

- For scenarios involving humans outside of the crash protection system of a vehicle (so-called "vulnerable road users," such as pedestrians and cyclists), safety is the top priority. Roadmanship could be ignored unless the action of the HAV may cause safety concerns for pedestrians, cyclists, or for vehicles behind or in adjacent lanes.

- Gap acceptance (the space between two vehicles when a driver decides to proceed) for an HAV making unprotected left turns and entering roundabouts are two examples where roadmanship should be considered.
- HAVs frequently engaging in hard braking (e.g., heavier than the 99.9 percent deceleration threshold, or 0.41g) for no human-apparent reason may also increase crash likelihood for following vehicles and should be penalized for lack of roadmanship.

Let's consider the unprotected left turn — there is no traffic light to signal the turn — as an example scenario to explain our concept of scoring HAVs for roadmanship. Assume there is an unprotected left turn in the target deployment area, and a sensor suite, such as a camera, is installed to observe how human drivers naturalistically interact with each other at that intersection. Once a large number of left turns have been collected, the data is analyzed. In particular, one variable that characterizes the challenge level of unprotected left turns is the "time margin to conflict zone" or Tcz. Imagine that at the intersection of the paths of the left-turning vehicle and the straight-driving vehicle there is a "conflict zone." The HAV that is making the left turn needs to make the following decision: Take the gap if it is large enough and turn, or wait for the straight-driving vehicle to clear the conflict zone first if the gap is short. Whether the gap is "large enough" or not can be determined based on what human drivers do. Figure 6 is obtained by the following process: The blue points indicate those time gaps, Tcz, that were rejected by human drivers, the red points indicate those time gaps, Tcz, that were accepted by human drivers. Because human drivers are not all the same, there are some overlaps. In the intermediate range, for example, a gap can be rejected by some drivers, but accepted by others.

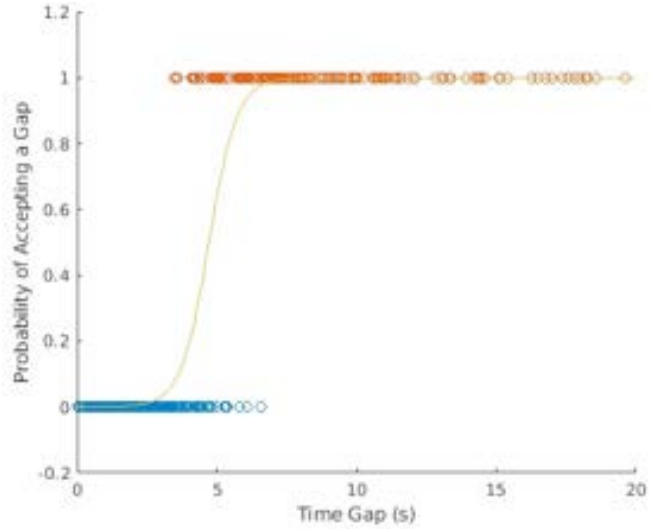

Figure 6. Rejected (blue) and accepted (red) time gap by human drivers in the unprotected left-turn scenario.

Figure 6 is based on actual left-turn data we collected from observing 2,000 left turns at a single intersection with a fixed speed limit. It can be seen that when a time gap is shorter than 3 seconds, all observed human drivers reject it. When the time gap is longer than 7.5 seconds, all observed human drivers take it. The acceptance rate gradually increases in between. Given this data, it is possible to design a 5-star rating system similar to the one shown in Figure 7. Again, all the numbers of Figure 7 should be treated as example ball-park numbers and not as mandates.

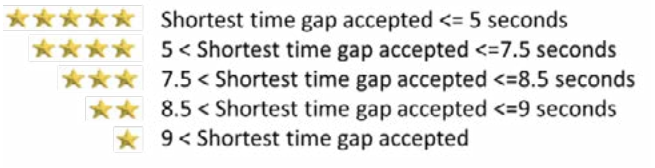

Figure 7. Example 5-star rating just for the unprotected left turn scenario.

If the results are intended for the general public, we recommend having just one pass/fail and one 5-star rating score for an HAV, instead of separate scores for each scenario

CONCLUSION

In order for society to accept HAVs as tools that are of value, safety must be demonstrated, and we must have a clear process that is open, rigorous, repeatable, and ever-improving. The Mcity ABC Test can contribute to the development of such a process. As envisioned, the Mcity ABC Test would provide a competency measure for an HAV in a controlled environment prior to performing additional on-road public testing, where the risk to human life is much greater. This paper has described the "B" portion of the test—behavior competence— and provides a set of methods by which these behavior competency tests could be carried out to measure the performance of a candidate HAV. By providing a set of tests randomly sampled from each required competency scenario, maintaining an equitable number of tests between candidates based on difficulty, and scoring the results as pass or fail for safety outcome, we have outlined a process we believe is open, rigorous, and repeatable. Gauging an HAV's roadmanship on a sliding scale adds an additional competency measure.

NEXT STEPS

These techniques are not intended to be unbending, but rather a spark for collaborative discussion, trial, and refinement.

Mcity intends to demonstrate the Mcity ABC Test with a neutral, noncommercial HAV platform we have developed here at the University of Michigan. If you would like to participate in this process, including using it to test your own HAV systems, please contact us at mcity@umich.edu.

About Mcity

Mcity at the University of Michigan is leading the transition to connected and automated vehicles. Home to world-renowned researchers, a one-of-a-kind test facility, and on-road deployments, Mcity brings together industry, government, and academia to improve transportation safety, sustainability, and accessibility for the benefit of society.

REFERENCES

1. https://www.ntsb.gov/Search1/pages/Results.aspx?k=HWY18MH010
2. https://dms.ntsb.gov/pubdms/search/document.cfm?docID=477717&docketID=62978&mkey=96894
3. https://www.nhtsa.gov/automated-driving-systems/voluntary-safety-self-assessment
4. Mcity ABC Test: A Concept to Assess the Safety of Highly Automated Vehicles
5. Laura Fraade-Blanar, Marjory S. Blumenthal, James M. Anderson, Nidhi Kalra, "Measuring Automated Vehicle Safety--Forging a Framework", RAND report, available at https://www.rand.org/pubs/research_reports/RR2662.html
6. Safety Pilot Model Deployment--Lessons Learned and Recommendations for Future Connected Vehicle Activities, FHWA-JPO-16-363, Sep 2015.
7. Integrated Vehicle-Based Safety Systems (IVBSS) Light Vehicle Field Operational Test Independent Evaluation, DOT HS 811 516, October 2011.
8. Lee, K., and H. Peng. "Evaluation of automotive forward collision warning and collision avoidance algorithms." Vehicle system dynamics 43.10 (2005): 735-751.
9. McGehee, Daniel V., Elizabeth N. Mazzae, and GH Scott Baldwin. "Driver reaction time in crash avoidance research: validation of a driving simulator study on a test track." Proceedings of the human factors and ergonomics society annual meeting. Vol. 44. No. 20. Sage CA: Los Angeles, CA: SAGE Publications, 2000.
10. Summala, Heikki. "Brake reaction times and driver behavior analysis." Transportation Human Factors 2.3 (2000): 217-226.
11. Zhao, Ding, et al. "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques." IEEE transactions on intelligent transportation systems 18.3 (2016): 595-607.