Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P0298-cd

Approach for a reliability proof of autonomous vehicles compared to human drivers through real-world road tests

Dr. Patrick Jäger

JHP GmbH, Germany. E-mail: patrick.jaeger@jhp-beratung.de

Dr. Melani Krolo

JHP GmbH, Germany. E-mail: melani.krolo@jhp-beratung.de

Ute Jäger

JHP GmbH, Germany. E-mail: ute.jaeger@jhp-beratung.de

Real-world testing, along with extensive simulation, is essential for validating autonomous vehicle performance under real operating conditions. However, proving the reliability of autonomous vehicles compared to human drivers through real-world road tests has been challenging due to the lack of a comparable database on human driving behavior within similar Operational Design Domains (ODDs). A study by UMTRI collected human ride hail data, including crash statistics, in San Francisco from 2016 to 2018, enabling a comparison between autonomous vehicles and human driving behavior over several years. The approach presented in this report evaluates the reliability of autonomous vehicles based on real-world tests on public roads, using data from Waymo autonomous vehicles operating as robotaxis in San Francisco. An analysis of disengagement data revealed improvements in later developmental stages for disengagements due to software discrepancies. The reliability of Waymo driverless vehicles in San Francisco was demonstrated and compared to human drivers from ride-hail services in the same ODD. Based on DMV data as of February 2024, it was shown that Waymo driverless vehicles in San Francisco are at least as reliable as human drivers with similar driving behavior in the same ODD. Additionally, even under conservative considerations, the reliability of human drivers was proven with a high confidence level.

Keywords: Reliability proof, autonomous vehicles, real-world testing, RGA (Reliability Growth Analysis), SRGM (Software Reliability Growth Model), disengagements, accidents with AV involved, comparable ODD.

1. Introduction and Overview

Before autonomous vehicles (AVs) can drive on our roads, they must be extensively tested and validated. The test validation includes all measures taken to ensure that the vehicle functions safely and reliably in every situation. This includes tests on the road, in simulation and in the laboratory. The on-road tests are particularly important as they reflect the real conditions under which the vehicle will operate. Various scenarios must be considered here.

To demonstrate the reliability of autonomous vehicles under real-world conditions, extensive mileage needs to be covered. To ensure that autonomous vehicles do not cause more collisions than human drivers, a comparison between driverless vehicles and human drivers must be conducted. One major difficulty in comprehending human driving capabilities lies in the fact, that various environmental conditions result in different crash rates. While comparing crash rates of autonomous vehicles and human drivers, it is essential to compare the driving behavior in comparable environments and to consider changing environmental conditions (e.g. temporary modifications, road construction, time of day, weather conditions, light and sight etc.).

Over the past several years, testing of autonomous vehicles on public roads has been conducted, accumulating a significant number

of miles driven. The California Department of Motor Vehicles (DMV) database contains valuable information regarding autonomous vehicles and their collisions and disengagements. This data provides insights into the performance and safety of autonomous vehicles in real-world driving conditions.

This paper introduces an approach on how the testing validation of autonomous vehicles on public roads can be carried out. The reliability of autonomous vehicles, that are being tested on public roads, are being compared to the human drivers within a matching ODD. A case study involving recently published data and statistics on human ride hail crash rates in San Francisco as well as DMV data of autonomous vehicle testing in San Francisco will further substantiate the approach in a practical manner. For the comparison between human drivers and autonomous vehicles, the crash rate with the primary contribution of driverless vehicles from a specific manufacturer is compared to a benchmark crash rate of a human driver with primary contribution. The approach and the statistical proof are presented in chapter 4.

Furthermore, disengagement data of the same autonomous vehicle fleet during the testing and operational phase of software and hardware development is statistically analyzed by using Reliability Growth Modeling (RGM) to evaluate and predict the reliability growth during testing of autonomous vehicles on the road.

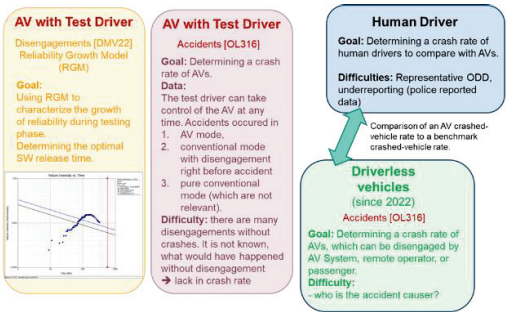

Fig. 1. An approach for a testing validation of autonomous vehicles on public roads

2. California Department of Motor Vehicles (DMV) Disengagement Data - Statistical Analysis Results

Manufacturers of autonomous vehicles participating in the Autonomous Vehicle Tester (AVT) Program and AVT Driverless Program must submit annual reports detailing the frequency of disengagements from autonomous mode during testing. These disengagements may occur due to technology failures or instances where the self-driving system cannot operate safely, requiring the test driver to take manual control. Factors contributing to this may include technical limitations, road conditions, or unforeseen events.

Given Waymo LLC's large vehicle fleet, which provides a solid foundation for statistical analysis, we have chosen to analyze their data further. After the published data [DMV22] of Waymo LLC vehicles in San Francisco has been viewed, we chose to set a filter regarding the starting dates (commissioning dates of the vehicles) beginning from 04/2019 – 11/2022 (11/2022: this was the most recent data at the time of the RGA analysis of autonomous vehicles with test driver), as from April 2019 on we could be sure, that the complete history of the chosen sample of Waymo LLC vehicles is known.

Figure 2 shows an overview of the different causes for disengagements and whether the test driver disengaged or the AV system. A detailed description of the different classes for causes is published in the California DMV Disengagement Reports.

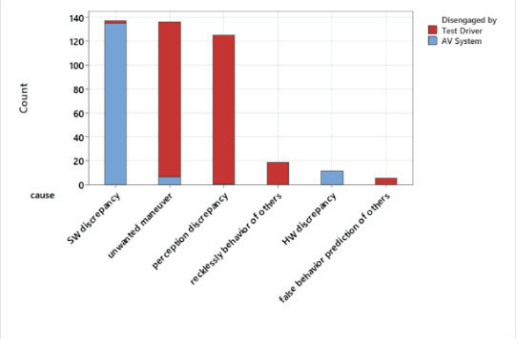

Fig. 2. Causes for disengagements for Waymo LLC vehicles 04/2019 – 11/2022 with test driver

Following chapter will show the RGA results for disengagements due to software (SW) discrepancy and perception discrepancy.

2.1. SRGM based on SW discrepancy

Software discrepancies have been observed exclusively between January 2021 and April 2022. The vehicles experiencing software discrepancies began operating between December 2020 and November 2021, except for one vehicle that began operating in December 2018. An analysis of the subset of vehicles commissioned during the period from December 2020 to November 2021 was conducted using ReliaSoft's Software RGA tool. The vehicle that started in December 2018 was not included in the analysis.

A reliability growth analysis was conducted on a specific subpopulation of Waymo LLC autonomous vehicles (with test driver) commissioned between December 2020 and November 2021, focusing on disengagements caused by software (SW) discrepancies. Additional information regarding SW versions unfortunately was not available. The majority of the SW discrepancies occur for vehicles within a narrow time frame starting in December 2020 and February 2021. Therefore, we assumed, that the vehicles were operating with the same software version and that SW updates (if there were any) took place for all the vehicles in the subpopulation.

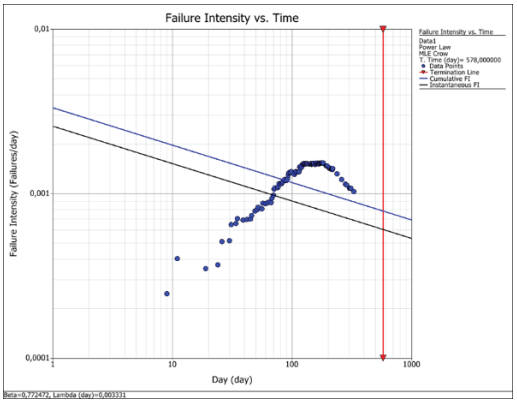

Fig. 3. Disengagement intensity over days due to SW discrepancy for subpopulation: 12/2020 – 11/2021

The analysis result shows that the data points of the disengagement intensity over days due to software discrepancies follow clearly an upside-down bathtub-shaped hazard function, also called a unimodal hazard rate function with a higher rate of failure during the middle of the expected lifetime. Such behavior can typically be observed during Software development. A unimodal hazard function is characterized by two shapes. Traditional models like the Power Law in ReliaSoft's Software Weibull++ are inadequate, as one can see in Figure 3, in describing this typical behavior.

Models like e.g. an inverse Weibull model or an inverse Lindley distribution, that are capable to represent an upside-down bathtub-shaped hazard function, are usually not integrated in standard reliability software tools. The RGM help to evaluate qualitatively the behavior of the disengagements over time and mileage and help to show the improvement over time and mileage. Any forecasts would need the establishment of a software tool that provides models that fit such data well. The model parameters would be obtained by the maximum likelihood estimation method. Such analyses will not be content of this paper. Qualitatively one can say, that for this sample there is a significant decrease in the disengagement intensity after approximately six months.

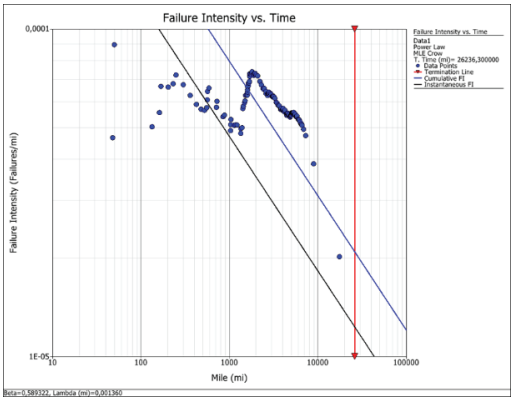

Fig. 4. Disengagement intensity over miles due to SW discrepancy for subpopulation: 12/2020 – 11/2021

Figure 4 shows the disengagement intensity vs. miles. Comparable to the disengagement intensity over days, there is a poor fit of the power law model to the disengagements due to SW discrepancies over miles as well. Therefore, forecasting based on this model is not possible. The intensity of disengagements follows a unimodal hazard rate for mileages greater than

1000. For mileages less than 1000, there is a zigzag pattern with fewer disengagements due to software discrepancies. If information about the vehicle's software versions and releases were available, a representative subpopulation could be defined for further analysis and prediction.

2.2. RGM based on perception discrepancy

A component of the vehicle's perception system (such as a camera, lidar, or radar) fails to detect objects accurately [DMV22]. Disengagements related to perception issues primarily occur in vehicles starting in December 2018 or in the year 2021. There are no recorded perception issues for vehicles starting between February 2022 and November 2022. Due to incomplete historical data for vehicles starting in 2018, a filter was applied to the RGA from April 2019 to November 2022 based on starting dates.

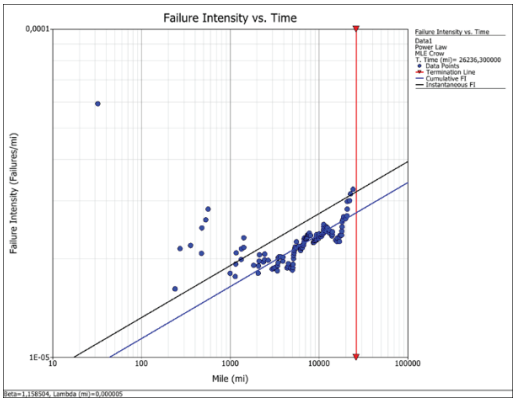

Fig. 5. Disengagement intensity over miles due to perception discrepancy 04/2019 – 11/2022

Reliability Growth Analyses have been accomplished for disengagements over days and over miles due to perception discrepancies for the Waymo LLC AVs with test driver for the starting time period 04/2019 – 11/2022. In contrast to the software discrepancy over days, with its typical upside-down bathtub-shaped hazard function and a significant improvement at the end, the perception discrepancy does not show the improvement in terms of a decreasing failure intensity, neither over days nor over mileage. After approximately 100 days, the second part of the upside-down bathtub is reached with an almost constant failure rate and a missing improvement at higher times and mileages. The quasi-constant failure rate after 100 days leads to an MTBF (here: mean time between disengagements) of approximately 4 years. The disengagement intensity over miles instead is increasing for higher mileages, especially for a mileage greater than 18.000 miles one can recognize a significant increase as shown in figure 5.

3. DMV autonomous vehicle accident data analysis

The Department of Motor Vehicles (DMV) in California maintains certain data related to autonomous vehicles (AVs) and collisions. It requires companies, that are testing AVs on public roads, to report any collisions involving their vehicles within 10 days. This includes collisions that result in property damage, bodily injury, or death. The reports of traffic collision involving an autonomous vehicle are published at https://www.dmv.ca.gov/portal and are freely accessible to the public. The “Report of traffic Collision Involving an Autonomous Vehicle” form OL 316, which can be viewed online, is used for reporting a traffic collision involving an autonomous vehicle.

Various steps were taken to analyze the collision data from the DMV California database. In this paper we summarize the relevant topics, that are being considered regarding the Waymo driverless vehicle reliability proof compared to the human driver.

In the subsequent evaluation, only collisions where Waymo driverless vehicles were identified as the cause were considered.

The collision reports were not fully documented, with a missing documentation of VINs in reports. In cases where it was not explicitly stated, whether a test driver was present, we assumed that there was no test driver and classified the collisions as involving driverless vehicles, what would lead to a more conservative calculation, as there might be an overrating of crashes for driverless vehicles.

21 crashes were recorded in the period of March 2022 – February 2024, assuming no test driver was on board. In most of the collisions, the AV was stopped in traffic before collision and the other party’s vehicle was moving and made contact with the AV. Out of the 21 crashes with a Waymo

driverless vehicle, five were defined by us as having the AV at fault. The determination was made to the best of our ability through studying the crash reports.

We have categorized the five remaining collisions [OL316] with the AV at fault regarding the classification into meaningful risk of injury and minor collisions. Three collisions were defined as minor collisions and two were defined as collisions with a meaningful risk of injury.

The accident categorization proved to be challenging due to insufficient or imprecise information (e.g. no velocity). Furthermore, it is not clear, what would have happened, if it was not e.g. the cardboard debris but an animal or human being instead. Would the AV interpret the situation correctly? We don't know. Due to these facts, we calculated several variants. We have defined three variants:

Table 1. Collision with driverless vehicle at fault and the categorization into minor crash or crash with meaningful risk of injury

Variants	Possible crashes* with meaningful risk of injury	Minor crashes
1	**2**	3
2	**1**	4
3	0	5

*From the total of five collisions that occurred, it can be deduced that in one case there is a possibility that this could be associated with a significant risk of injury. In another case, this possibility is somewhat less likely, but still exists.

4. Human driver vs. autonomous vehicle

The majority of car crashes are caused by human error, such as distracted driving, speeding, drunk driving, or fatigue. Autonomous vehicles, on the other hand, are not prone to these human errors, as they rely on sensors, algorithms, and artificial intelligence to make decisions. While autonomous vehicles have the potential to significantly reduce crashes caused by human error, they are not immune to all collisions. For example, technical malfunctions, software bugs, unforeseen scenarios can lead to crashes involving autonomous vehicles. The technology continues to advance and improve and it must be extensively tested and validated to assure that the overall crash statistics for autonomous vehicles will be lower compared to human drivers.

The key question is, how many miles does an autonomous vehicle have to drive collision free in order to prove that it is at least as reliable as the human driver?

To answer the question, different data has been analyzed to reflect the current field behavior of human drivers and autonomous vehicles within a matching ODD and to evaluate the on-road test progress of a fleet of autonomous vehicles in San Francisco, California.

4.1. Human Ride hail Crash Rate

The University of Michigan Transportation Research Institute (UMTRI) published a study "*Establishing a Crash Rate Benchmark Using Large-Scale Naturalistic Human Ridehail Data*" in a whitepaper in September 2023, UMTRI-2023-18.pdf (umich.edu).

The research involved collecting human ride hail data over a two-year period from 2016 to 2018 through a collaboration between Cruise, General Motors (GM), UMTRI, and the Virginia Tech Transportation Institute (VTTI). The Operational Design Domain (ODD) for the study included the entire city of San Francisco (SF), with the exception of certain high-speed roads (e.g., those with posted speeds exceeding 35 mph). The UMTRI fleet and VTTI fleet collectively covered a total of 5.611.765 ODD miles during this investigation.

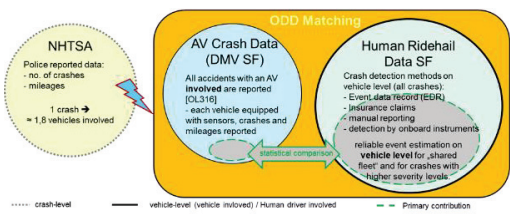

Fig. 6. Comparison of the reliability of human-driven ride hail vehicles vs. driverless vehicles in SF

The human ride hail crash rate for the shared fleet is one crash in 15.414,4 ODD driving miles [UMTRI23].

The whitepaper does not include information about a meaningful risk of injury or primary contribution. Cruise published a follow-on blog post with further detail on injury outcomes [Cruise23]. Crashes with meaningful risk of injury have been defined as crashes that measure Level 2 or more (L2+). There were 60 L2+ crashes observed in the UMTRI Fleet and six L2+ crashes observed in the VTTI fleet, which produced an estimate of 66 L2+ ODD crashes in a total of 5,611,763 ODD miles. This results in a human ride hail crash rate with meaningful risk of injury (L2+) in the observed ODD of one crash in 85,027 miles; or, 11.76 L2+ ODD crashes per million miles [Cruise23].

Corresponding to the fact, that usually one party is generally found to be primarily responsible for a given two-vehicle crash [Cruise23], the mileage per crash (all crashes) has been doubled to get the crash rate on vehicle-level with primary contribution.

If all crashes are considered without a specific classification regarding meaningful risk of injury, on vehicle-level with primary contribution this would result in a human ride hail crash rate for the shared fleet of:

one crash in 30.828,8 ODD driving miles

The human ride hail L2+ crash rate on crash-level is: **one L2+ crash in 85.027 ODD driving miles.**

What does this statistic mean for an on-road test plan for autonomous vehicles?

Let us discuss this in the following chapter.

4.2. Test plan for AVs based on chi-squared distribution for constant failure rate

How many miles does the driverless vehicle fleet need to drive in the same ODD (San Francisco on roads with max. speed 35 mph) without a crash to prove the human ride hail MTBF (including all crashes) with a confidence level of $P_A = 95\%$?

The equation to calculate the scope of testing that is required to prove a minimum life for exponentially distributed failure behavior with a defined confidence level is given in Eq.(1):

$$\sum_{i=1}^{n} t_i = \frac{1}{2} MTBF_{\min} \cdot \chi^2_{2(x+1);C} \quad (1$$

t_i = mileage of each vehicle i
x = number of failures
n = sample size
X^2 = value of chi-square distribution
α = probability of error
C = confidence level

Figure 7 shows the result of the test plan for AVs in the same ODD as the human ride hail vehicles. This test plan is calculated for the vehicle-level. Therefore, the human ride hail mileage per crash on crash-level, which is published in [UMTRI23], is doubled in order to get the MTBFs on vehicle-level with primary contribution. The test plan is calculated for each fleet, the UMTRI fleet, the VTTI fleet and the shared fleet.

Based on the shared fleet with a human ride hail MTBF of 30.828,8 ODD miles with primary contribution, driverless vehicles would need to drive 92.355 miles crash free in the same ODD.

And a conservative approach (as not all crashes are being recorded in the UMTRI fleet, which leads to a higher MTBF): Based on the UMTRI fleet with a human MTBF of 55.294 ODD miles with primary contribution, driverless vehicles would need to drive 165.646 miles crash free in the same ODD. Please note, that this testing scope resulting from a) is probably overly sharp.

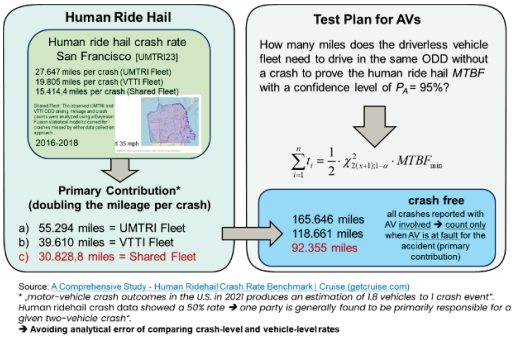

Fig. 7. Test plan for AVs based on human ride hail crash rates

But what about crashes with meaningful risk of injury? The testing scope discussed above

does not consider any severities. To ensure that autonomous vehicles do not contribute to an increase in collisions resulting in injuries compared to human drivers, extensive testing must be observed. The following example serves to illustrate this point based on the human ride hail statistic in SF with risk of injury. As already discussed above, the human ride hail L2+ crash rate on crash-level is one L2+ crash in 85.027 ODD driving miles. Therefore, the human ride hail MTBF on vehicle-level with primary contribution is 170.054 miles.

How many miles does the driverless vehicle fleet need to drive in the same ODD without a crash to prove the human ride hail MTBF with meaningful risk of injury (L2+) with a confidence level of P_A = 95%? Fig. 8 shows the result of the test plan.

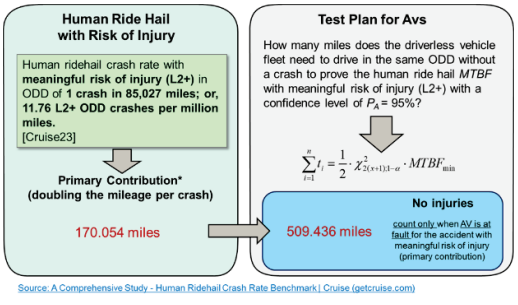

Fig. 8. Test plan for AVs with regard to a risk of injury

In San Francisco, driverless vehicles would need to travel approximately 510 thousand ODD miles without causing any injuries to surpass the reliability record of human ride-hail drivers in terms of crash statistics with a meaningful risk of injury. The determination of a testing scope regarding fatal collisions was not possible, as information regarding the number of fatal collisions of human drivers was missing. The following chapter will present the reliability proof that the current driverless vehicle fleet of the manufacturer Waymo LLC, operating in San Francisco, has demonstrated for driving behavior up to February 2024.

4.3. Statistical proof

The reliability is demonstrated, based on the chi-squared distribution for constant failure rates, for the Waymo driverless vehicles without test driver present. There were 14 disengagements documented for the driverless fleet (as of February 2024), which will not be counted as collisions. The result will represent the reliability proof for the current development stage of the Waymo driverless fleet with the possibility to be disengaged by a remote operator. The total mileage of the Waymo driverless vehicles is 1.240.569 miles (data base from disengagement reports [DMV22] as of 02/2024). Mileages for 18 Waymo LLC driverless vehicles were recorded in the disengagement reports from March 2022 to November 2022 [DMV22]. In total, there are records of 124 driverless vehicles with mileages documented in 2022 and 2023 (as of February 2024).

The statistical proof is calculated for all variants. The 14 disengagements still took place and are not considered as any possible crashes. Variant 1 represents the conservative approach with two collisions categorized as possible crashes with meaningful risk of injury. The MTBF of 170.054 miles has been proven with a confidence level of 97,6%. Variant 2 includes one collision categorized as a possible crash with meaningful risk of injury. The MTBF of 170.054 has been proven with a confidence level of 99,4%. Variant 3 is the categorization with 5 minor crashes (zero meaningful risks of injury crashes). The MTBF of 30.828,8 has been proven with a confidence level of > 99,9999%.

If one would presume, that the 14 disengagements would have ended in 14 minor crashes with AV at fault, the reliability proof for these 14 “possible crashes” + 5 minor crashes is: The MTBF of 30.828,8 miles has been proven with a confidence level of PA > 99,98%. For the proposed variants one can say based on the given DMV data as of February 2024, that the Waymo driverless vehicles achieved the reliability goal in terms of the human ride hail mean time between crashes with very high confidence.

5 Summary of the Statistical Analysis

Until now, with today’s autonomous vehicle fleets, it was impossible to give a statistical prove, that autonomous vehicles are at least as reliable as human drivers based on real-world road tests. For the comparison, a comparable database about the human driving behavior within matching ODD was missing. Based on recently published data by

DMV and UMTRI, the presented approach in this report to evaluating the reliability of autonomous vehicles compared to human drivers based on real-world tests on public roads has been applied in a practical manner. Waymo autonomous vehicles operating as robotaxis in San Francisco have been examined regarding disengagement data and crash data. It has been demonstrated how disengagement data of Waymo AVs can be used to monitor the failure / disengagement intensity over time and how it is possible to track an improvement over the developmental phase. Furthermore, it was shown, based on the given DMV data as of February 2024, that the Waymo driverless vehicles operating in the specific area in San Francisco as robotaxis are at least as reliable as the human driver with similar driving behavior in the same ODD with a very high probability of confidence. Even for the conservative consideration of two driverless vehicle collisions categorized as possible crashes with meaningful risk of injury, the human MTBF of 170.054 has also been proven with a high confidence level of 97,6%. Nevertheless, crash rates can vary even within a well-defined ODD as different driving environments can vary regarding the risk to drivers (time of day, time of year, traffic density, fog, light etc.). Rigorous simulations must be conducted to maximize the safety of AVs while minimizing potential risks on our roads.

6 Discussion Section: The Challenges of AVs

The introduction of AVs has the potential to improve road safety, enhance mobility for individuals and reduce traffic congestion; yet it also raises significant concerns that require careful consideration. AVs are increasingly dependent on complex software systems to navigate and make real-time decisions. This makes them vulnerable to software glitches in terms of potential bit-flips, which can lead to unexpected behavior or malfunctions while on the road. This issue can be handled practically by applying modern safety related chip architectures and functions like the lock step procedure. Additionally, the connectivity features that enable communication with other vehicles and infrastructure also expose AVs to potential cyber-attacks. Malicious actors could take advantage of these vulnerabilities and take control or disrupt operations. As the technology continues to grow, addressing these concerns is crucial for the safety of AVs. Public and legal resistance to AVs often arises from fear of the unknown, particularly concerning safety due to high-profile accidents. To address this skepticism, transparency in testing is essential, along with community engagement to clarify AV technology and build trust. Legal frameworks must also adapt to new challenges, such as developing insurance models that account for shared liability between manufacturers and users. Additionally, the absence of human judgment in AV decision-making raises ethical dilemmas, like e.g. how to prioritize passenger safety versus that of pedestrians or prioritizing passenger safety over property damage in unavoidable accident scenarios. AVs operate based on pre-programmed algorithms that may not account for complex moral dilemmas. This raises ethical questions about how AVs should be programmed to respond in emergency situations. The lack of human intuition and empathy in these scenarios underscores the importance of integrating ethical frameworks into AV design and operation. Accessibility is another important concern. AVs have the potential to improve transportation options for people who currently have limited access, such as those in underserved communities. However, if AVs are not introduced carefully, there is a risk that they could increase existing inequalities instead of helping to reduce them. In summary, addressing these challenges is essential to ensure a safe, ethical, and equitable integration of AVs into our transportation systems.

References

Zhang, L. (2023). Human ridehail crash rate benchmark. Blog post, [Cruise23].

DMV (2024). State of California, Department of Motor Vehicles: Disengagement Reports. [DMV22]

DMV (2024). OL 316, Report of traffic Collision Involving an Autonomous Vehicle (www.dmv.ca.gov).

Carol Flannagan*, Andrew Leslie*, Raymond Kiefer†, Scott Bogard*, Geoff Chi-Johnston‡, Laura Freeman‡, Rayman Huang‡, David Walsh‡, Antony Joseph‡ (2023). Establishing a Crash Rate Benchmark Using Large-Scale Naturalistic Human Ridehail Data. * University of Michigan Transportation Research Institute (UMTRI); † General Motors LLC; ‡ Cruise LLC . [UMTRI23]

John M. Scanlon, Kristofer D. Kusano, Laura A. Fraade-Blanar, Timothy L. McMurry, Yin-Hsiu Chen, Trent Victor: Benchmarks for Retrospective Auto-mated Driving System Crash Rate Analysis Using Police-Reported Crash Data. 2023, Waymo LLC.