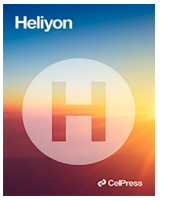
Heliyon 10 (2024) e34379

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Comparative safety performance of autonomous- and human drivers: A real-world case study of the Waymo Driver

Luigi Di Lillo [a,b,c,d,*], Tilia Gode [e,f], Xilin Zhou [a], Margherita Atzei [a], Ruoshu Chen [e,f], Trent Victor [e]

[a] Swiss Reinsurance Company, Ltd, Switzerland
[b] Autonomous Systems Laboratory, Stanford University, USA
[c] Frazzoli Group, ETH Zürich, Switzerland
[d] Zardini Lab, MIT, USA
[e] Waymo LLC, USA
[f] Casualty Actuarial Society, USA

A B S T R A C T

After several years of public road testing, the commercial deployment of fully autonomous vehicles—or Automated Driving Systems (ADS)—is poised to scale substantially following significant technological advancements and recent regulatory approvals. However, the fundamental question of whether an ADS is safer than its human counterparts remain largely unsolved due to several challenges in establishing an appropriate real-world safety comparison method. As scaling ensues, the lack of an established method can contribute to misinterpretations or uncertainties regarding ADS safety and impede the continuous and consistent assessment of ADS performance. This study introduces three research developments to define a robust and replicable safety comparison method to address this critical methodological gap. First, we introduce the use of liability insurance claims data to measure the comparative safety between ADS and human drivers. Second, we use Swiss Re insurance claims data to establish the first zip code- and responsibility-calibrated human performance benchmark, composed of over 600,000 private passenger vehicle claims and 125 billion miles of driving exposure. Third, we perform a case study by applying the developed baseline to evaluate the safety impact of the Waymo Driver. We find that when benchmarked against zip code-calibrated human baselines, the Waymo Driver significantly improves safety towards other road users. The comparison method established in this study can be replicated for other regions or ADS deployments to aid the decision-making of ADS safety stakeholders such as regulators, and instill trust in the general public.

1. Introduction

As the deployment of Automated Driving Systems (ADS) continues to scale, a fundamental question is whether ADS is safer than human drivers. However, drawing scientifically sound conclusions about the safety of ADS versus human drivers has long faced largely unsolved methodological challenges. Valid safety comparisons must select an appropriate safety performance metric, overcome differences in collision reporting standards between ADS and human-driven vehicles (HDV), account for the underreporting of human driving collisions, and set a safety-relevant collision severity threshold. Furthermore, they should use human driver data specific to the region in which the ADS operates while maintaining robustness, and apply a statistical method to measure the uncertainty of the results [1–6]. Improperly controlled variations across collision datasets can lead to inflated or deflated collision statistics.

Safety comparison research should also consider crash responsibility, as the level of contribution to a crash is an important factor in assessing a driver's safety performance [7]. A responsibility-calibrated comparison requires a human baseline that reports the number

* Corresponding author. Swiss Reinsurance Company, Ltd, Switzerland.
E-mail addresses: Luigi_DiLillo@swissre.com, luidilillo@gmail.com (L. Di Lillo).

https://doi.org/10.1016/j.heliyon.2024.e34379
Received 25 April 2024; Received in revised form 14 June 2024; Accepted 9 July 2024
Available online 14 July 2024

of instances (i.e., frequency) in which a human driver contributes to a safety-relevant incident. In addition, it must circumvent inconsistent crash causation reporting requirements across police jurisdictions and the informal nature of crash causation assignments in police reports [8,9]. Responsibility-calibrated comparisons can sharpen safety insights and improve precision and transparency when communicating with the public and regulators. However, real-world safety impact studies are yet to include responsibility-calibrated comparisons.

Well-designed safety comparisons can shape evidence-based public policies, inform insurance risk assessments, increase public transparency, and set future safety standards for ADS deployments. Additionally, establishing scientifically robust safety comparison methods is critical for an accurate, continuous, and consistent assessment of the performance of ADS deployments as scaling ensues. Producing such methods and safety comparisons will require careful consideration of the methodological challenges at hand.

2. Framework

This study introduces the use of real-world *auto third-party liability insurance claims data* (henceforth referred to as liability claims data), to measure and compare ADS and HDV safety performances. A liability claim is a request for compensation when someone is responsible for the damage to property or injury to another person, typically following a collision. In this study, we do not investigate events which do not result in a liability claim—such as non-contact events like "near misses"—which may offer interesting insights into ADS safety but are outside the scope of our data and, thus, of the current work.

Liability claims data are an alternative to existing performance measures, such as collisions reported to the police, and naturalistic driving data collected using on-vehicle sensors in real-world driving conditions. Claims data are also uniquely suited to addressing the current challenges facing ADS-human driver comparison research, due to several advantages:

Consistency. Reporting standards for liability claims are consistent across ADS and human drivers. Unlike police-reported collisions which have varying severity thresholds across ADS and human drivers [3,6], liability claims are traditionally registered for any motor incident resulting in property damage or bodily injury, regardless of the vehicle's level of automation.

Comprehensiveness. Liability insurance claims offer a more comprehensive assessment of human driving incidents than collision databases from police reports because (a) claims data have a higher reporting frequency of lower-severity collisions or injuries in human drivers [1,10], and (b) police reports do not capture non-collision-related injuries [11] and capture fewer instances of injury claims [10] (see Fig. 1).

Safety-Relevance. Unlike police-reported crashes that underreport lower-severity collisions and naturalistic driving data that report any contact event, insurance data capture motor incidents that result in harm or the need for financial recovery, including lower-severity collisions and injuries, but excluding minor events that produce no damage or harm. This severity threshold is recognized as a uniquely intuitive and straightforward measure of safety-relevant crashes in safety impact research [10].

Robustness. Widespread auto insurance requirements in the US mean that insurance data exist for a vast majority (~87 %) of drivers across all states.[1] As a result, insurance data can generate statistically robust human performance benchmarks that can be calibrated consistently and robustly for different driving regions without the need for supplementary human driving data collection.

Responsibility-Calibration. Liability claims data uniquely capture information regarding crash or injury causation contributions[2] because collision responsibility is directly determined during the liability claims adjudication process [12]. Liability claims adjudication, reporting, and their development are designed using insurance industry best practices to assess crash causation contribution and predict future crash contributions.

Finally, liability claims data are the core intellectual property of insurance and reinsurance companies, and are the key ingredient for underwriting, costing, and pricing risk. They provide a snapshot of the risk landscape within a given industry (e.g., personal transportation) and a glimpse into an insurer's underwriting strategy. As a result, they are highly sought-after within the insurance industry as well as the industries it insures.

3. Case study

This study uses two insurance claims datasets provided by Swiss Reinsurance Company Ltd. and Waymo LLC to produce a responsibility-calibrated human performance benchmark for ADS safety comparison research, and to apply this benchmark to assess the safety performance of the Waymo Driver. Specifically, the case study examines how the safety performance of the Waymo One™ ride-hailing service in San Francisco, CA and Phoenix, AZ compares to that of human drivers within the same regions, as measured by the frequency of real-world liability claims. The Waymo Driver, the core of the Waymo One service, is a level 4 Automated Driving

[1] Date (insurance-research.org).

[2] After a collision, the percentage of responsibility assigned to each party is determined by an insurance claims adjuster. Liability in the insurance context is distinct from the legal concept of fault. For the purposes of this study, we conservatively assume that the policyholder contributed to the collision giving rise to injuries or third-party property damage if the claim was resolved with a liability payment by means of the insurance claims adjustment process. Also, for the purposes of this study, we are using claims that are resolved, or likely to resolve with a liability payment, as a proxy for partial or full responsibility.

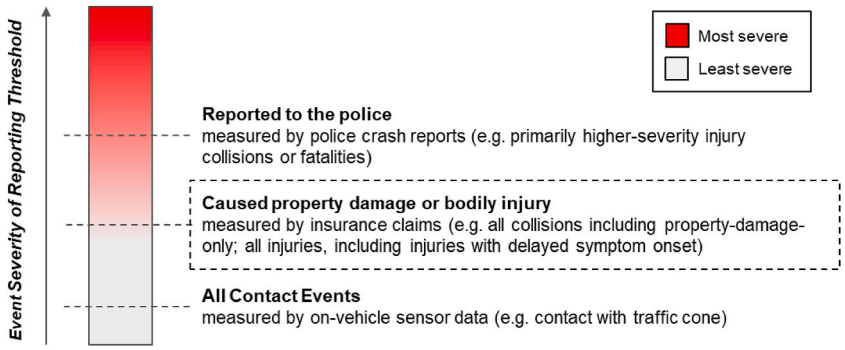


Fig. 1. Comparison of insights carried by different datasets. The "All Contact Events" set includes events which are not safety-relevant. The "Reported to the police" set, on the other hand, is biased toward high-severity events and lacks standardization in reporting.

System (ADS) as defined in SAE J3016 (SAE 2021) and does not require a human driver behind the wheel when in Rider-Only (RO), "driverless" operation.[3]

We analyze claims filed under third-party liability (see footnote 2) insurance policies that drivers are required to carry by law in California and Arizona,[4] split by Property Damage Liability and Bodily Injury Liability coverages. Property Damage Liability insures against damages that at-fault drivers cause to other people's vehicles and property, typically during crashes. Bodily Injury Liability insures against medical, hospital, and other expenses for injuries that at-fault drivers inflict on the occupants of involved vehicles or others on the road.

4. Methods

The Waymo data included in this study are from miles driven and claims that occurred and were reported during the period from January 1, 2018, to August 1,2023.[5] Waymo's claims and mileage are organized into four categories based on driving modes: manual,[6] testing operations (TO) which are monitored by an autonomous specialist (also known as a "safety driver") in the driver's seat,[7] rider-only (RO) in which no human is behind the steering wheel,[8] and TO + RO[9] which includes all ADS (non-manual) miles.

The human driver baselines are based on Swiss Re's property damage liability (PD) and bodily injury liability (BI) claims data for claims that occurred between 2016 and2021[10]. For this study 600,000 claims and over 125 billion miles of exposure were used (i.e., post-zip code and responsibility calibration filtering), summed across over 25 million vehicle insurance policy years.[11] Since insurance claims data include information on cost severity rather than injury severity, we do not differentiate between levels of injury severity. In addition, as one collision can lead to both bodily injury and property damage claims, such collisions are included in both the BI and PD

[3] cpuc-av-program-applications-guidance-20211026.pdf The CPUC application process makes the distinction between: The "Drivered AV Passenger Service" program allows for the provision of passenger service in test AVs with a driver in the vehicle; The "Driverless AV Passenger Service" program allows for the provision of passenger service in test AVs without a driver in the vehicle.

[4] Auto Insurance Requirements - California DMV.

[5] Claim count development is considered by reviewing known events. Future development (additional claim counts) is still possible from unreported and/or underreported claims. For Waymo, claims emergence months after the collision date is less likely due to the ability to detect event occurrence in a timely manner compared to human driven vehicles.

[6] "Manual" is a mode in which the Waymo vehicle is driven manually (i.e., without the ADS engaged) by an autonomous specialist (human driver) who can react to a dynamic environment and operating vehicles under strict safety guidelines.

[7] Testing operations ("TO") is a phase of public road testing in which the ADS is engaged under monitoring of a trained autonomous specialist (human driver) who is seated in the driver's seat and can take over the driving task at any time. In "TO" mode, the ADS is engaged to operate the vehicle under monitoring of trained human autonomous specialists. A collision is categorized under TO as long as the ADS was engaged at any time during the 5 s leading up to the impact. A collision could still be categorized under TO under the 5-s rule when a human maneuver may have led to the collision, leading to a conservative estimate.

[8] Rider-only ("RO") is a mode where the ADS operates the vehicle without any human behind the steering wheel.

[9] Testing operation and Rider-only ("TO + RO") is the combination of the TO and RO datasets.

[10] Both our human and ADS analyses include six years of the most reliable and up-to-date data. In actuarial science, due to the length of time it takes to report, process, and close a claim, it is best practice to consider claims data from a policy year to be stable 1–2 years after a policy year ends to account for claims reported and/or maturing outside of the policy year. This practice is not as relevant for Waymo's data due to the ability to detect event occurrence in a timely manner (see footnote 5). As a result, our human performance benchmark covers 2016–2021, while our Waymo ADS data covers 2018–2023. Per-million-miles human driving fatality statistics for the US demonstrate a greater rate of traffic deaths in the years 2022 and 2023, than 2016 and 2017. As a result, the exclusion of 2022 and 2023 and inclusion of 2016 and 2017 may have led to a baseline which may be more conservative (lower rate) than a year-range-matched baseline. NHTSA Crash Stats: Early Estimate of Motor Vehicle Traffic Fatalities in 2023 (dot.gov).

[11] A policy year is a 12-month period of insurance coverage, beginning from the effective date of the inception of the insurance policy.

comparisons.

The baseline was calibrated using both mileage (driving exposure) and zip codes (geographic region). For the mileage- and zip code-calibrated baselines, claims associated with vehicles registered to addresses (i.e., where the insured resides) within Waymo's operating zip codes in San Francisco and the Phoenix metropolitan region were included. Although the use of the zip code of the registered vehicle address, as opposed to the collision, is a limitation of this study, we expect this to have a small-to-negligible impact on the frequency estimate. We further elaborate on the topic of zip codes in the **Discussion of Case Study** section.

Since Waymo's claims exposure is measured using mileage, to produce a valid human baseline for comparison, we convert the number of exposure *years* contained in the HDV dataset to the number of exposure *miles*. To do so, we estimate the annual vehicle miles traveled (VMT) per vehicle.[12] The annual VMT per vehicle is estimated at the yearly and regional (state or city) granularity.

Our estimation method uses aggregate mileage data to calculate the average mileage per vehicle. For mileage data, we use VMT statistics provided by the Federal Highway Administration (FHWA), which reports the total monthly VMT by state and total annual VMT by urbanized area[13] [13,14]. These statistics are based on individual state reports of traffic data counts collected using permanent automatic traffic recorders on public roadways. For per-vehicle statistics, we use the FHWA's annual vehicle registration statistics and US Census Data [15,16]. For both San Francisco and Phoenix and for each of the six coverage years included in the baseline (2016–2021), we produce two VMT per vehicle estimates: one estimate that draws from state VMT data and another estimate based on VMT data per urbanized area. For the analysis presented above, we chose to use estimates based on state VMT data because they yielded a lower (more conservative) baseline frequency, given that the state average annual miles per vehicle are higher than those of the urbanized areas.[14] For further elaboration on mileage estimation choices and the predicted impact on the results, please refer to the **Discussion of Case Study** section.

For the mileage- and zip code-calibrated baselines, claim frequencies are independently calculated for vehicles within the San Francisco and Phoenix metropolitan areas. To build baselines for comparison, these frequencies are weighted according to Waymo's mileage distribution across Phoenix and San Francisco. Since the different Waymo mileage categories (RO, TO, TO + RO, and Manual) have different Phoenix-San Francisco mileage distributions and operating zip codes, separate HDV baselines are calculated for each of the four mileage categories.

For significance testing, we conclude that there is a statistically significant difference between claims frequencies if the 95 % confidence intervals of Waymo's operations and their corresponding human baselines do not overlap. That is, that there is a difference even when accounting for statistical uncertainty due to the small mileage volume. For the human baseline, due to the large sample size, we use a normal approximation confidence interval and take the mileage distribution between San Francisco and Phoenix into account when computing the standard errors. For Waymo's operations, due to the smaller sample and exposure sizes, confidence intervals are calculated using the Poisson Exact Method [17],.[15]

5. Results

Fig. 2 shows the results of the comparison of the bodily injury and property damage liability insurance claims for the Swiss Re human driver baselines and Waymo data for the manual (A), testing operations (B), rider-only (C), and testing operations + rider-only (D) mileage categories.

When Waymo vehicles were driven fully **manually** for 14,436,298 miles for data collection, bodily injury claims frequency was reduced by **45 %** compared with the Swiss Re human driver baseline (0.55 vs 1.01 claims per million miles), but an overlap in 95 % confidence intervals ($Manual_{BI}$ *95 % CI* [0.24, 1.09], *Baseline 95 % CI* [1.00, 1.02]) indicates that there is insignificant or inconclusive evidence of this reduction. Property damage claims frequency was significantly reduced by **34 %** (2.22 vs 3.34 claims per million miles). This result was confirmed by non-overlapping confidence intervals ($Manual_{PDL}$ *95 % CI* [1.52, 3.13], *Baseline 95 % CI* [3.33, 3.36]). These results are displayed in Fig. 2A.

While driving 35,228,320 miles in **testing operations (TO)** mode, the Waymo Driver, together with autonomous specialists (also known as "safety drivers"),[16] significantly reduced bodily injury claims frequency by **92 %** (0.09 vs 1.09 claims per million miles), TO_{BI} *95 % CI* [0.02, 0.25], *Baseline 95 % CI* [1.08, 1.09]. Property damage claims frequency was significantly reduced by **95 %** (0.17 vs 3.17 claims per million miles), TO_{PD} *95 % CI* [0.06, 0.37], *Baseline 95 % CI* [3.16, 3.18]. These results are displayed in Fig. 2B.

While driving without a human behind the steering wheel in **RO** mode for 3,868,506 miles, the Waymo Driver reduced bodily

[12] A generally accepted estimate for annual VMT per vehicle in the US is approximately 11,000 miles. However, because of variations in driving patterns across different US cities and states, we separately estimate annual VMT per vehicle for each region (city or state) within the baseline. Alternative Fuels Data Center: Maps and Data - Average Annual Vehicle Miles Traveled by Major Vehicle Category (energy.gov).

[13] An urbanized area is defined as an area with 50,000 persons that at a minimum encompasses the land area delineated as the urbanized area by the U.S. Census Bureau.

[14] For the Phoenix Metropolitan Area, the estimated Arizona average annual VMT per vehicle is 1.01× higher than that of the Phoenix urban area across the years of 2016–2021. For San Francisco, the estimated California average annual VMT per vehicle is 1.18× higher than that of the San Francisco area.

[15] Note that the baseline's confidence intervals are created using estimated annual mileages, and different estimation assumptions will yield different point estimates and confidence intervals.

[16] Waymo One is assessed as a service which includes both the Waymo Driver safety performance as well as the autonomous specialist safety performance while in TO phase, as the claims outcomes are a product of both.

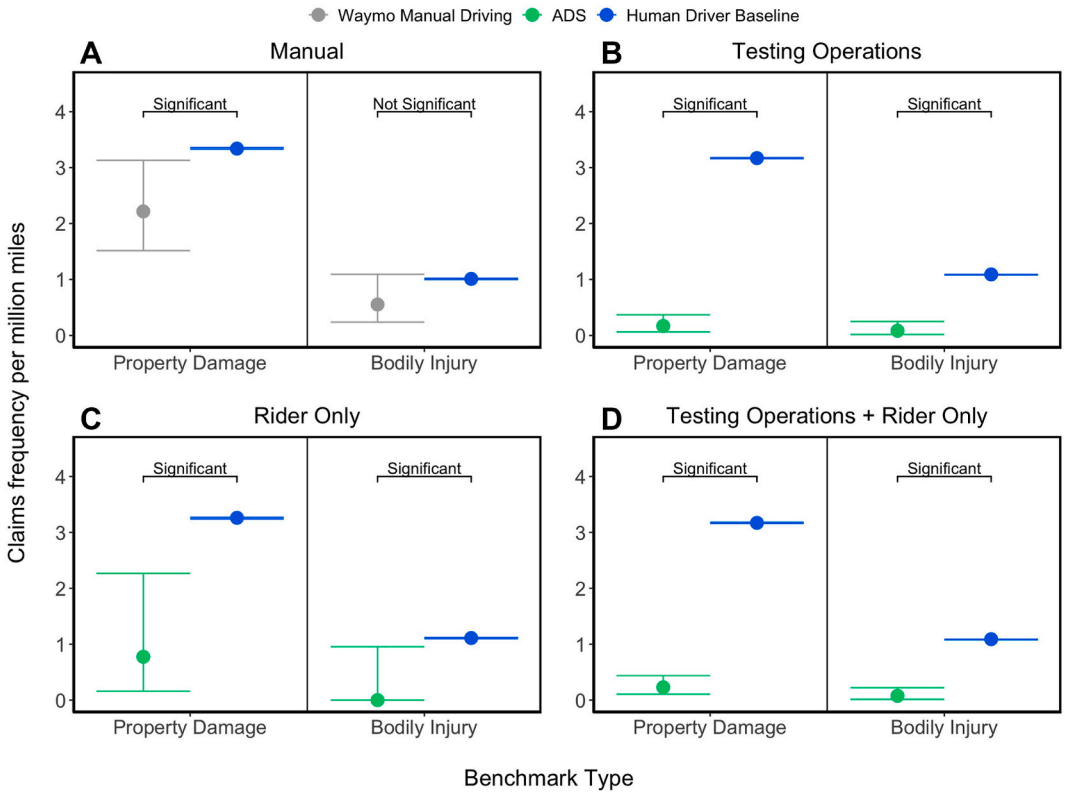


Fig. 2. Comparison of Swiss Re human driver baselines with Waymo liability insurance claims for Manual (A), Testing Operations (B), Rider-Only (C), and Testing Operations + Rider-Only (D) for property damage and bodily injury. Significant = non-overlapping 95 % CIs; Non-significant = overlapping 95 % CIs or inconclusive results. ADS=Waymo One Automated Driving System, Baseline = Swiss Re private passenger vehicle (human driver) baselines, calibrated for each mileage category.

injury claims frequency by **100 %**, or zero claims, (0.00 vs 1.11 claims per million miles). The difference is statistically significant, as indicated by the non-overlapping confidence intervals (RO_{BI} *95 % CI* [0.000, 0.95], *Baseline 95 % CI* [1.10, 1.12]). This provides evidence of the ADS' ability to reduce bodily injuries on public roads. Property damage claims frequency was significantly reduced by **76 %** (0.78 vs 3.26 claims per million miles), as indicated by non-overlapping 95 % CIs (RO_{PD} *95 % CI* [0.16, 2.27], *Baseline 95 % CI* [3.24, 3.27]). These results are displayed in Fig. 2C.

When **TO and RO** datasets were *combined*, totaling 39,096,826 miles, there was a significant reduction in bodily injury claims frequency by **93 %** (0.08 vs 1.09 claims per million miles), $TO + RO_{BI}$ *95 % CI* [0.02, 0.22], *Baseline 95 % CI* [1.08, 1.09]. Property damage claims frequency was significantly reduced by **93 %** (0.23 vs 3.17 claims per million miles), $TO + RO_{PDL}$ *95 % CI* [0.11, 0.44], *Baseline 95 % CI* [3.16, 3.18]. These results are displayed in Fig. 2D.

6. Discussion of case study

In this study, the baseline for comparison was derived from a human population of insured drivers residing in the same zip codes as the Waymo's service. The benefits of this population include its size and robustness, which lends itself to narrower confidence intervals. In addition, the selected baseline population is likely the population that may use Waymo services instead of driving themselves. In future studies, we plan to investigate other populations and methods to subset the data and generate additional comparative insights.

A limitation of the selected human baseline is that the locations of crashes that generate claims are unknown, limiting the ability to filter claims based on Waymo's Operational Design Domain (ODD). As a result, whereas the Waymo ODD largely does not include freeway driving in the dataset, the human database includes miles driven and claims that might have occurred on freeways. Due to variations in the collision frequency per million miles between freeways and non-freeways, this may have led to a baseline that is more conservative (lower rate) than a roadway-matched baseline. In Scanlon et al. [6], it was highlighted that the surface street passenger vehicle crash rate was 18 % higher than the all-road crash rate, which includes freeway driving. We would expect a similar relationship to hold for claim count.

In addition, the human benchmark comprises claims associated with vehicles registered to addresses (i.e., where the insured resides) within Waymo's operating zip codes. As a result, claims that involve an HDV garaged in the selected zip codes but occur outside

of the selected region may also be included in the study. Previous studies have shown that most injury collisions occur within a small radius from residency [18], and that American drivers rarely travel far from their place of residence, with approximately 80 % of one-way household trips being less than 10 miles.[17] Given the comparably much larger aggregate square mileage of the case study's target geographical regions,[18] only a small number of claims included in the human benchmarks are expected to originate from collisions occurring outside of the target region. As a result, the impact of the use of zip codes of residence (as opposed to of collision) on the human benchmarks is expected to be small to negligible. Moreover, for most insurers, the only insureds' contextual variable captured and used for costing and pricing the policy is the zip code, and it is common practice in actuarial science and insurance ratemaking that the garaging zip code is representative of a zip code's risk exposure. Finally, with the increase in telematics and automated vehicles, we expect that zip-code-based events will gain relevance in insurance datasets over time.

In addition, the annual per-vehicle average mileage used in the case study is an estimate based on aggregate mileage and population data. This estimated average mileage naturally exhibits a degree of estimation uncertainty that can impact the point estimates of the human benchmark. For this estimation uncertainty to result in a "false positive," that is, a significant demonstration of positive safety impact which otherwise would be found inconclusive, our human driving mileage estimates would have to have been *underestimated* (which would lead to an overestimated baseline). Given the high degree of mileage underestimation required for a false positive to occur within the TO and RO + TO mileage categories, any estimation uncertainty should only reasonably affect the RO and Manual comparison results. As a precaution against such false positives, we chose to use mileage estimates which would yield a lower baseline frequency (i.e., choosing state-based aggregate data over urbanized-region-based aggregate data). As a result, we expect our estimation to yield a conservative baseline. Future related work can decrease the level of estimation uncertainty by incorporating insurance telematics data that precisely track vehicle mileage, or vehicle mileage data based on reported odometer readings.

Finally, although the primary objective of this work is to establish a robust baseline to compare liability claims rates for ADS and HDV without accounting for causality effects (i.e., what led an ADS or HDV into a collision), as ADS operations scale up, ADS - HDV interactions could potentially have an impact on ADS and HDV claims frequencies. In their recent work, Zheng et al. [19] investigated the impact of driving style and type of driving interaction on drivers' decision-making and subjective feelings in mixed traffic situations. Given the relatively much smaller number of ADS compared to HDV in this study and the lack of drivers' emotional descriptions in the claims' reports, it stands to reason that any consideration about the causality effects due to mixed traffic situations would be weak.

7. Closing remarks

This study introduces the use of private passenger vehicle (human driver) liability insurance claims data to establish performance baselines to benchmark the safety performance of an ADS. Our evidence suggests that the Waymo One service is significantly safer (i.e., regarding the percentage of reduction in the number of liability claims) than the zip code- and responsibility-calibrated private passenger vehicle baselines established by Swiss Re, corresponding to over 600,000 claims and over 125 billion miles of exposure, and calibrated to match Waymo's mileage distribution across operating locations.

The introduced methodology overcomes many of the existing challenges facing ADS and human driver crash rate comparisons and can be applied within the industry to assess the safety performance of additional ADS deployments. This study also highlights the importance of cross-industry research collaborations to co-develop expertise and knowledge for public benefit, such as ensuring the safe advancement and deployment of technology and evaluating and informing the public about the technology's impact on societal well-being.

Data availability statement

This manuscript uses insurance liability claims data to i) build a statistically robust baseline representative of human driver performance and ii) argue and draw conclusions on the safety performance of ADS. As described in the text, liability claims data offer substantial advantages over accident statistics data and naturalistic data. However, they are the key and core assets of any insurance company, which uses them to run retrospective analyses, price policies, and segment and steer their portfolios. Due to the commercial sensitivity of these datasets, the liability claims data supporting this study cannot be made publicly available.

The code and data that support the mileage conversion will be made openly available, along with its documentation of use. These data were derived from resources available in the public domain published by the Federal Highway Administration's Office of Highway Policy Information and the U.S. Census Bureau.

Significance statement

Are autonomous drivers safer than their human counterparts? For almost a decade, the autonomous vehicle—or Automated Driving System (ADS)—industry has indicated that ADS adoption will lead to significant improvements in road safety. However, several

[17] FOTW #1230, March 21, 2022: More than Half of all Daily Trips Were Less than Three Miles in 2021 | Department of Energy; FOTW #1042, August 13, 2018: In 2017 Nearly 60 % of All Vehicle Trips Were Less than Six Miles | Department of Energy.

[18] The human driving baselines used to benchmark Waymo's ADS mileage categories cover over 1300 square miles. Benchmarks for Waymo's RO mileage cover over 400 square miles (San Francisco region = 48 square miles; Phoenix region = 379 square miles).

methodological challenges have made this assertion difficult to verify. As ADS deployments prepare to scale, the general public and regulators need the necessary insights and statistics to make informed decisions regarding road safety and to assess the safety impact of ADS-related policy decisions. This study tackles this long-standing challenge by establishing an empirical methodology and a case study comparing ADS and human driver performance using unique insurance data.

CRediT authorship contribution statement

Luigi Di Lillo: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Tilia Gode:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Xilin Zhou:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Margherita Atzei:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Ruoshu Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Trent Victor:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Verena Brufatto, Solutions Data Analyst, and **Federica Capparelli**, Data Scientist, have been instrumental in creating the baselines and analytics around this work. Binbin Li, Risk Analyst, has contributed greatly to putting together the statistical analysis and ensuring data quality. **Ross Amend**, Claims Expert, was instrumental in ensuring data accuracy, **Matt Glascock**, Motor Pricing Expert, has offered invaluable support by sharing his expertise on motor claims data and by reviewing the methodological assumptions and the analytical work. **Andrea Biancheri**, Senior Pricing Actuary, and **Cristiano Misani**, Lead Advanced Scoring, have been critical in challenging the analysis and the results.

References

[1] L. Blincoe, T. Miller, J.-S. Wang, D. Swedler, T. Coughlin, B. Lawrence, F. Guo, S. Klauer, T. Dingus, The Economic and Societal Impact of Motor Vehicle Crashes, 2019 (Revised) (Report No. DOT HS 813 403), National Highway Traffic Safety Administration, 2023, February.
[2] T. Victor, K. Kusano, T. Gode, R. Chen, M. Schwall, Safety performance of the Waymo rider-only automated driving system at one million miles. www.waymo.com/safety, 2023.
[3] M. Blanco, J. Atwood, S.M. Russell, T.E. Trimble, J.A. McClafferty, M.A. Perez, Automated Vehicle Crash Rate Comparison Using Naturalistic Data, Virginia Tech Transportation Institute, 2016.
[4] M. Lindman, I. Isaksson-Hellman, J. Strandroth, Basic numbers needed to understand the traffic safety effect of automated cars, in: IRCOBI Conference, 2017, September, pp. 1–12.
[5] J. Bärgman, M. Svärd, S. Lundell, E. Hartelius, Methodological challenges of scenario generation validation: a rear-end crash-causation model for virtual safety assessment, Transport. Res. Part F Traffic Psychol. Behav. (2024) 374–410.
[6] J.M. Scanlon, K.D. Kusano, L.A. Fraade-Blanar, T.L. McMurry, Y.H. Chen, T. Victor, Benchmarks for retrospective automated driving system crash rate analysis using police-reported crash data, arXiv preprint, https://arxiv.org/pdf/2312.13228, 2023.
[7] S. Klauer, T. Dingus, V. Neale, J. Sudweeks, D. Ramsey, Comparing Real-World Behaviors of Drivers with High versus Low Rates of Crashes and Near Crashes. (Report No. DOT HS 811 091), National Highway Traffic Safety Administration, 2009.
[8] K. Kim, I. Brunner, E. Yamashita, Modeling fault among accident—involved pedestrians and motorists in Hawaii, Accid. Anal. Prev. 40 (6) (2008) 2043–2049.
[9] A. Williams, V. Shabanova, Responsibility of drivers, by age and gender, for motor-vehicle crash deaths, J. Saf. Res. 34 (5) (2003) 527–531.
[10] I. Isaksson-Hellman, M. Lindman, An evaluation of the real-world safety effect of a lane change driver support system and characteristics of lane change crashes based on insurance claims data, Traffic Inj. Prev. (2018) 104–111.
[11] B. Mills, J. Andrey, D. Hambly, Analysis of precipitation-related motor vehicle collision and injury risk using insurance and police record information for Winnipeg, Canada, J. Saf. Res. (2011) 383–390.
[12] E.R. Braver, R.E. Trempel, Are older drivers actually at higher risk of involvement in collisions resulting in deaths or non-fatal injuries among their passengers and other road users? Inj. Prev. : journal of the International Society for Child and Adolescent Injury Prevention 10 (1) (2004) 27–32.
[13] Office of Highway Policy Information, Highway Statistics Series (2023). https://www.fhwa.dot.gov/policyinformation/statistics/2020/hm71.cfm.
[14] Office of Highway Policy Information, Traffic Volume Trends (2023). https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm.
[15] Office of Highway Policy Information, State motor-vehicle registrations. https://www.fhwa.dot.gov/policyinformation/statistics/2021/mv1.cfm, 2023.
[16] U.S. Census Bureau. ACS Demographic and Housing Estimates. https://data.census.gov/.
[17] F. Garwood, Fiducial limits for the Poisson distribution, Biometrika 28 (3/4) (1936) 437–442.
[18] B. Haas, A.G. Doumouras, D. Gomez, C. de Mestral, D.M. Boyes, L. Morrison, A.B. Nathens, Close to home: an analysis of the relationship between location of residence and location of injury, J. Trauma Acute Care Surg. 78 (4) (2015) 860–865.
[19] Ma Zheng, Z. Yiqi, Driver-automated vehicle interaction in mixed traffic: types of interaction and drivers' driving styles, Hum. Factors 66 (2) (2024) 544–561.