Safety Performance of the Waymo Rider-Only Automated Driving System at One Million Miles

Trent Victor, Kristofer Kusano, Tilia Gode, Ruoshu Chen, Matthew Schwall

Waymo, LLC

Abstract

This paper examines the safety performance of the Waymo Driver™, Waymo's Automated Driving System (ADS). It analyzes one million miles of driving on public roads in parts of California and Arizona with no human behind the wheel – what we call rider-only (RO) operations. There were no reported injuries, and only two collisions that were comparable to the National Highway Traffic Safety Administration's Crash Investigation Sampling System ("CISS"), a nationally representative collision database of collisions that were police reported and in which at least one vehicle was towed. There were an additional 18 minor-contact events that were too minor to meet the tow-away and police-report criteria for CISS, where nine of these 20 contact events had no damage. Neither the CISS-comparable- nor the minor-contact events were intersection related or involved VRUs (Vulnerable Road Users), and every vehicle-to-vehicle event involved one or more road rule violations and/or dangerous behaviors on the part of the other vehicle's operator. 55% of all contact events occurred with a stationary Waymo vehicle, and 40% were parking related, with some overlaps between the two categories. These results, in combination with previous reported data and past studies performed by Waymo on fatalities reconstruction and on collision-avoidance testing, support the assertion that the Waymo Driver is successful at reducing injuries and fatalities today. To establish a valid comparative assessment of the Waymo Driver, it is most efficient to apply the same requirements to ADS and Human comparison, only counting events that are likely included in human-reported crash data benchmark, for example, only CISS-compatible collisions. Confidence in conclusions about the safety of the Waymo Driver will increase continuously as both credibility of available data and the validity of predictions improves over time.

Introduction

This paper builds on previous work (Schwall et al., 2020; Scanlon et al., 2020, 2021; Kusano et al., 2022) examining the real-world safety performance of the Waymo Driver™, an Automated Driving System (ADS) that is capable of performing the entire dynamic-driving task on a sustained basis without the need for human intervention. The Waymo Driver is a level 4 ADS, as defined in SAE J3016 (SAE 2021) and does not require a human driver behind the wheel. The Waymo Driver is deployed in the Waymo One™ ride-hailing service, which currently operates in parts of California and Arizona. The service includes operations without a driver behind the wheel, referred to as rider-only (RO) operation, and is available to transport members of the general public.

Many different stakeholders are interested in the safety of ADSs, including the general public, policymakers, regulators, standards organizations, the insurance industry, and the research community. Decades of scientific research and understanding across multiple domains, from locomotive and automotive to aviation, in development and standardization, and across public, private, and military sectors, have all informed the current definition of *Safety* as *absence of unreasonable risk (AUR)* (ISO 26262:2018, International Standards Organization [ISO], 2018)[1]. These definitions culminate in safety acceptance criteria representing the absence of unreasonable risk (ISO, 2022). Such acceptance criteria are predicated upon safety performance indicators that can sit at any point along the causal chain that connects various forms of activation mechanisms to the potential manifestation of harm for the identified hazards. While there is still an ongoing discussion regarding what constitutes an acceptance criterion or "valid societal moral concepts" for residual risk targets (for probability of occurrence of harm and the severity of that harm), Waymo's safety philosophy is to reduce traffic injuries and fatalities, that is, to achieve what we call a Positive Safety Impact (PSI)[2].

As detailed in Webb et al. (2020), Waymo's process for making safety readiness determinations entails an ordered examination of the relevant metrics and outputs from a number of safety methodologies combined with careful safety and engineering judgment focused on the specific facts relevant for a particular determination. Waymo approves a specific configuration of the ADS for a given deployment when it determines the ADS is ready to operate in its Operational Design Domain (ODD). In Waymo's readiness determinations (see Webb et al., 2020), prospective safety impact estimation methods are used[3]. These include re-simulating a certain software version and system configuration against historical logs and predicting potential collisions on public road driving with counterfactual simulations (Schwall et al., 2020). For analysis of reconstructions of fatalities, counterfactual simulations have also been used (Scanlon et al., 2021, 2022).

---

[1] Risk is defined as the combination of the probability of occurrence of harm and the severity of that harm (ISO, 2018). In turn, Unreasonable Risk is defined as risk judged to be unacceptable in a certain context according to valid societal moral concepts (ISO, 2018). Acceptance Criteria is defined as a criterion representing the absence of an unreasonable level of risk (ISO, 2022).
[2] The reader may be familiar with the external term of Positive Risk Balance (see Favarò, 2021). However, the notion of PRB and its relation to the determination of AUR is still being shaped externally (Favarò, 2021), so Waymo prefers to use the explicit call out to a positive contribution to safety impact, as defined by the extensive traffic safety literature on safety benefits (see Scanlon, et al., 2021).
[3] *Prospective* safety impact methods aim to estimate the effect of a safety system before it is widely deployed. The prospective safety impact methodology approach aims to predict future effects, as opposed to a *retrospective* study, which observes the benefits of a safety system after it has been widely deployed in the field (Scanlon et al., 2021).

Retrospective studies of real-world collision outcome data are important for many reasons. Waymo believes strongly in transparency, and this one million rider-only mile achievement represents both a significant milestone and an opportunity to continue a dialogue, started in Schwall et al. (2020), on issues associated with outcome assessment and reporting. One million rider-only miles is equivalent to over 80 years of driving for the average U.S. licensed driver (McGuckin & Fucci, 2018). Although one million rider-only miles is a large amount of data, we acknowledge the limitations of this amount of data to capture rare high-severity situations of high injury potential. Retrospective studies of safety outcomes need sufficient accumulation of exposure (e.g., miles and/or hours of operation) to confidently draw conclusions. High confidence in lagging indicators, such as injurious outcomes, will be achieved when sufficient exposure allows for statistical rigor in retrospective, a-posteriori measurement. This paper focuses on interpretation of actual outcomes observed over one million rider-only miles.

There are two ways in which a new technology, like an ADS, can potentially reduce injury risk compared to a benchmark, as shown in Figure 1 (Korner, 1989; Kullgren, 2008; Lindman et al. 2010). The first way is to reduce the severity of the harmful events (shown as (a) in Figure 1). In (a), the total frequency of events is not changed, but the crash severity is reduced. The second way to reduce risk is to reduce the overall frequency of harmful events, while the relative distribution of crash severity is held constant (shown as (b) in Figure 1). An effective ADS does both (a) and (b) simultaneously, resulting in a distribution that moves both downward and to the left.

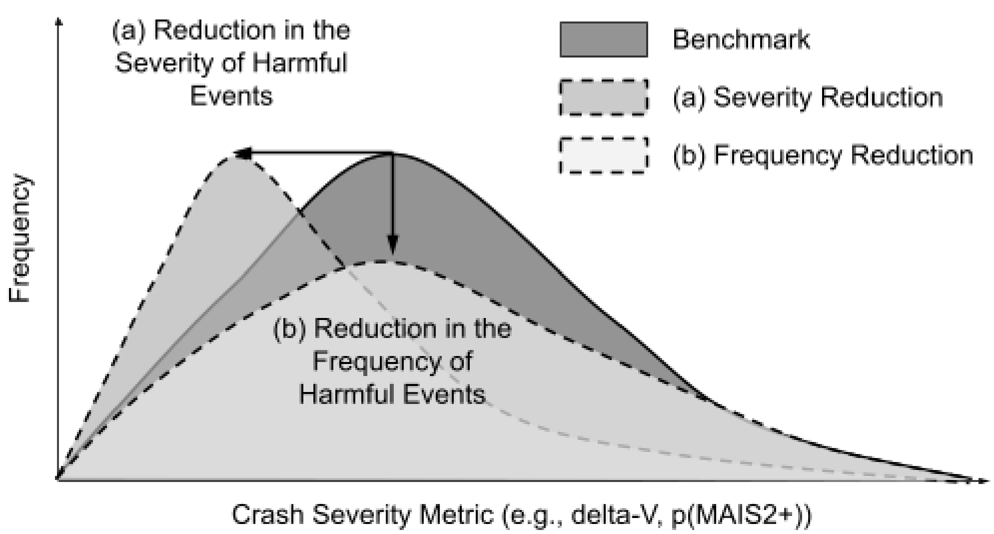


Figure 1. Illustration of reduction in injury risk through reduction in severity and frequency of harmful events.

Evaluation of collision severity

Injury risk in automotive collisions can be evaluated using dose-response models that link the dose (e.g., the impact severity measured by change in velocity during the crash) to the response (the injury outcome) (Kullgren et al., 2010). The reason such models are useful is that occurrence of an injury in a contact event can be affected by many factors, including the age and physical condition of the vehicle occupants and/or the type of vehicles and their safety features. For example, in crashes with similar impact speeds, older occupants are more likely to be injured than younger occupants (McMurry et al., 2021) and occupants in cars with 5-star crash safety ratings are less likely to be injured than occupants in lower rated vehicles (Kullgren et al., 2010). To be able to compare the severity of different contact events objectively, a risk of injury is often used by traffic safety researchers instead of relying on observed actual injury outcomes in order to control for these different factors that may contribute to injury outcome and are different between different crash populations. This injury risk formulation is especially used in prospective safety benefit analyses when performing counterfactual simulations (Scanlon, et al., 2021).

Evaluation of collision frequency

The concept of determining the frequency of collisions, either in the overall count of collisions, or a rate of collisions per mile driven or hour operated, is an easy-to-understand concept. An occurrence rate (e.g., number of crashes per mile) based on ADS' on-road driving performance data can be derived. Such estimates can come with wide confidence intervals, especially when they are established on relatively low mileage. As more miles are accumulated, the confidence interval becomes narrower, leading to a more "trustworthy" central estimate. This is known in common actuarial practice in the insurance sector as the concept of *Credibility* (see Appendix 1), which represents the degree of believability one can place on observed empirical data when trying to predict the future (Mahler & Dean, 2001). However, fully conclusive evidence in the determination of safety of a new technology like an ADS depends on (a) availability of retrospective, a-posteriori, safety outcomes and/or (b) the validity of predictions of such outcomes. However, excellent safety performance can create a sort of "credibility irony," where improved performance leads to rarer observations of safety outcomes and larger requirements for mileage to gain sufficient statistical confidence. As explained in Appendix 1, through comparison with a robust human baseline, detection of safer performance can be done with fewer collision observations, and therefore lower mileage. The requirement for a robust human benchmark points to the need to appropriately calibrate or adjust the available human data to ensure a valid comparison with an ADS. Although it has been estimated that an ADS may need to be driven for hundreds of millions or billions of miles to determine if a rate of fatal collision is less than that of humans (Kalra & Paddock, 2016; Lindman et al., 2017), this is still a topic of some debate (Young, 2021). The credibility discussion in Appendix 1 implies that a safer ADS can require less mileage if a human benchmark is used. Further development of a suitable benchmark is needed.

Calibration of human benchmark data to ADS data is essential to enable proper comparison and has many challenges including these key issues:

- *Representative high-severity collisions*. Benchmark data must have enough exposure (mileage) to contain a representative sample for rare collisions (serious injuries and fatalities, unusual conflict types). Nationally representative collision databases, such as

CISS and FARS, contain reliable high-severity collision sampling for most but not all conflict types (e.g. CISS does not contain VRUs, heavy vehicles), and it can be challenging to find representative ODD-specific datasets. Naturalistic Driving Study (NDS) datasets, despite having collected data for over 50 million miles, still lack enough mileage to contain high severity collisions (serious injuries and fatalities) (Blanco, et al., 2016).

- *Correction of under reporting of collisions.* The benchmark and ADS data must be calibrated to correct for collision identification and under reporting issues. For example, 24.3% of injury collisions and 59.7% of property-damage-only collisions are not reported to police in CISS (Blincoe et al., 2015). CISS does not correct for under reporting. ADSs likely have a higher likelihood of identification and reporting of collision events due to sophisticated sensing and fleet operations management. Calibration is needed for consistency in likelihood of identification and reporting in benchmark and ADS data sets because collision rates can be lower simply because of low reporting likelihood, and be falsely interpreted as safer.
- *ODD-specific matching.* The frequency of collisions within conflict types is highly dependent on context (VRU density, traffic density, weather, etc), as seen in regional variations in collision databases (Cai et al., 2017; Schwall et al., 2020). ADS and benchmark data must be matched on ODD. This leads to a situation termed the *ODD dilemma*, the tradeoff between using large, robust datasets (e.g., nationally representative databases) and small, ODD-specific datasets (Scanlon et al., 2021). Note that ODD matching differences between ADS and Naturalistic Driving Studies (NDS) datasets also exist. Similarly, misinterpretations can occur when assessing traffic risks simply as a ratio of total fatalities to total travel distances (Redelmeier, 2014) as this may obscure important factors such as regional variations.
- *Vehicle characteristics* like type, age, size and fitment of advanced driver assistance systems may also need to be accounted for (Lindman et al., 2017).

On the safety significance of minor and severe events

As discussed above, aggregate data from crash reporting will not be available to demonstrate a retrospective positive safety impact of an ADS with high degree of confidence for some time. However, crash reports from ADS-equipped vehicles may be useful as a diagnostic for early detection of potential safety issues, were one to exist, and organizations such as NHTSA and the California DMV have adopted mandatory reporting requirements[4,5]. The events presented in

[4] In the U.S., the National Highway Traffic Safety Administration (NHTSA) issued a Standing General Order (SGO) on June 29, 2021 that, among other requirements, requires reporting of certain “crash” events involving an ADS, where a “crash” is defined as “any physical impact between a vehicle and another road user (vehicle, pedestrian, cyclist, etc.) or property that results or allegedly results in any property damage, injury, or fatality” (Carlson, 2021). The California Department of Motor Vehicles (DMV) requires manufacturers of autonomous vehicles operating under an AV testing permit to report any “collision originating from the operation of the autonomous vehicle on a public road that resulted in the damage of property or in bodily injury or death” (California Code of Regulations, 2014).
[5] The diagnostic power of crash reporting improves with increased exposure. For this reason, national reporting systems are expected to be more sensitive and able to detect concerning trends more quickly

this paper include data not only from collisions that involve an injury and/or property damage, but also from more minor *contact* events that did not result in either injury or property damage. Non-damage contact events are included in this paper for transparency and to investigate whether these events have value as surrogates (leading indicators) of potentially more safety-relevant events.

The idea of minor, non-injurious events serving as a surrogate for more severe events originated in the study of workplace safety as Heinrich's Triangle, see Figure 2 (Heinrich, 1950; Chang et al., 2017). In the field of traffic safety, the research area of surrogate measures of safety encompasses all the measures of safety that do not rely on crash data and instead are meant to be an alternative or a complement to analyses based on crash records (Chang et al., 2017). The first surrogate measures of traffic safety were based on the observation of near-collisions and traffic conflicts in the 1950-60s in Traffic Conflict Theory, and this field has been active as an alternative to the recognized shortcomings of relying on crash data, particularly their rarity. However, a generalizable method to correlate surrogate measures with crash data is still lacking and is an area of research (Arun et al., 2021).

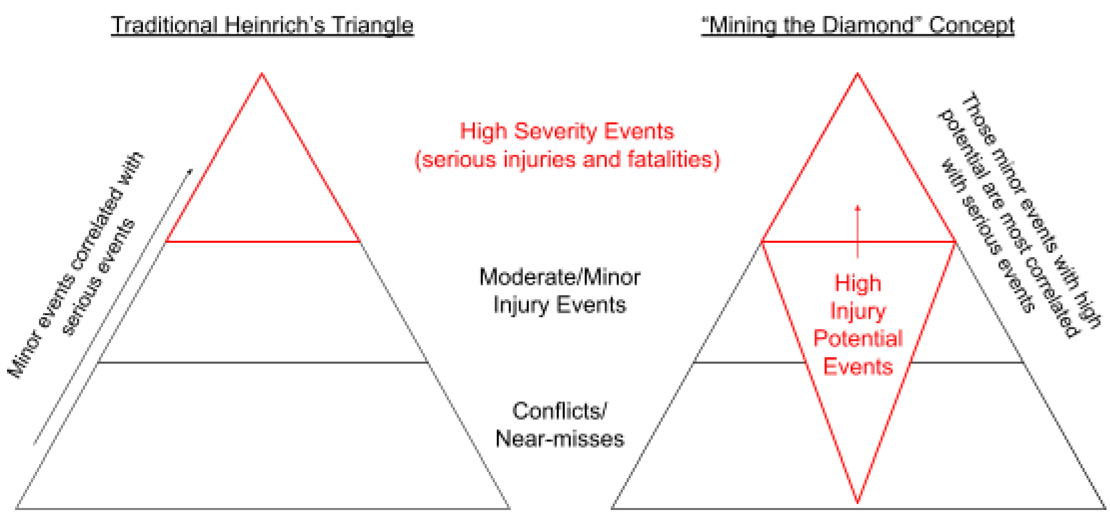


Figure 2. Illustration of Using Minor Injury Events and Near-misses with High Injury Potential to Predict High Severity Events (i.e., "Mining the Diamond" concept from Smith & Jones, 2013).

Heinrich's Triangle is based on the concept that more frequently-occurring near misses and/or minor events can be correlated to the rate of serious injury and fatal events. The reasoning behind Heinrich's triangle is that many minor or non-injury events may have similar causal mechanisms to serious injury or fatal events, and evidence supports that workplaces that reduce minor events do experience less fatalities (Yorio & Moore, 2018). Even still, the contributing factors present in some high severity events are often not present in the frequently occurring near misses (Manuele, 2011). This insight led to a modification of Heinrich's Triangle, that others termed "mining the diamond" (Smith & Jones, 2013), which is depicted in Figure 2.

than state or local reporting systems, especially as ADS operators are starting to test in multiple locations simultaneously.

This “mining the diamond” concept acknowledges that many minor events aren’t indicators of more severe events and, instead, focus should be put on preventing those events that have the potential of resulting in higher severity outcomes. The premise is that better correlation between surrogate metrics and serious outcomes might be achievable by considering only those surrogate events that could potentially result in a serious outcome.

Support for such a dissociated relationship between high and low severity events in traffic collisions – and therefore rebuttal of Heinrich’s Triangle – can be seen in collision data from before and during the COVID pandemic. Compared to prior years, during the pandemic there was a reduction of low severity collisions yet an increase in fatalities (Grembek et al., 2022; Meyer, 2020). The potential to dissociate low and high severity collision rates is reflected in Vision Zero, which focuses on severe injury and fatal collisions (Abebe et al., 2022). In Vision Zero, a road safety measure that leads to an overall decline in fatalities and serious injuries is preferable, even if it involves a greater number of collisions or minor injuries (Abebe et al., 2022). A “mining the diamond” approach would be supported over Heinrich’s triangle if some minor events (conflicts or contact events) contain the necessary factors to be able to potentially escalate to higher severity, but other minor events do not. The contact events shared in this paper are intended to provide a full dataset to examine whether non-damage contact events are indicative of higher severity collisions.

Aim

The aim of this paper is (a) to examine all contact events experienced during the first one million miles of rider-only (RO) operations of the Waymo Driver, an ADS, and (b) explore what conclusions can be made from this observed real-world safety performance in terms of the frequency and severity of these contact events. In the methods section, we will describe how we operationalized the definitions of the contact events reported in this study, how we estimated injury risk for these contact events, and which human crash datasets we will use to compare to these ADS contact events. In the results section we will provide descriptions (dates, conflict types, collision partners, vehicle damage, and narrative description) of the identified contact events. Finally, we end this paper with discussion and conclusions.

Methods

Waymo contact events

The data presented in this paper is from Waymo's ride hailing operations on public roads without a human behind the steering wheel, also known as rider-only (RO) mode. The miles were collected from Waymo's fully autonomous ride hailing service, which operated in parts of California and Arizona at all hours of the day and night. The current operational design domain (ODD) in these locations generally includes roadways with speed limits up to and including 45 miles per hour and excludes highways and inclement weather such as heavy rain, heavy fog, and dust storms.

Waymo's rider-only ride-hailing operations reached its first one million rider-only miles on January 21, 2023, on an AV fleet using both the Jaguar I-PACE and the Chrysler Pacifica platforms.[6] Rider-only ride-hailing operations began in 2019 in the East Valley region of the Phoenix metro area using the Chrysler Pacifica platform, followed by San Francisco in March 2022 and the area around downtown Phoenix in May 2022 using the Jaguar I-PACE platform. In this paper, we will refer to the service operating in the Phoenix East Valley region on the Chrysler Pacifica platform as the "PHX 4th Generation" operations. We will refer to the service operation in San Francisco and downtown Phoenix using the Jaguar I-PACE platform as either "SF 5th Generation" or "PHX 5th Generation."

This paper reports on all contact events, defined as any physical, energy transferring contact between the ADS-equipped vehicle and another object or human. This includes events that range from low-speed contact with a construction pylon to collisions with another vehicle. We excluded contacts with soft road debris such as a plastic bag or empty cardboard box and where only the tire of the ADS-equipped vehicle contacts a part of the road environment, such as a curb, resulting in no visible vehicle damage. We also excluded contact events where a person (a pedestrian, cyclists, or motorcyclist) intentionally touched the Waymo vehicle. These types of intentional touching events can occur often in dense urban environments where people are passing around stopped vehicles.

For each contact event, a description of the damage, if any, is provided for the Waymo AV and the other vehicle or object involved. Damage to the Waymo AV was determined by inspection, and damage to other vehicles or objects was determined by inspection and/or as reported by the vehicle owner. Road wear, such as scuffs and light scratches that could be polished out, is not included in the damage descriptions.

Calculation of injury risk in the Waymo dataset

For this study, we defined the risk of injury to occupants within the vehicle in terms of a maximum Abbreviated Injury Scale of 2 or greater (MAIS2+) (Association for the Advancement of Automotive Medicine, 2016). Examples of AIS2 level injuries are concussions with no or brief loss of conscience, fractures to the sternum, and 2 or few rib fractures. An impulse momentum collision model was applied to the impact conditions as detected by the ADS vehicle sensors to estimate the change in velocity (delta-V) and principal direction of force (PDOF) as described in Scanlon et al. (2021). Inertial properties of the Waymo equipped vehicle were measured. Inertial properties of the collision partner were estimated based on the size of the object as perceived by the ADS vehicle sensors. The delta-V and PDOF were used in combination with population average values for belt use, occupant age, and vehicle model year in an omni-direction injury

[6] Waymo first operated an ADS-equipped vehicle on public roads without a human driver in Austin, Texas in 2015 using a specialized vehicle without a steering wheel or accelerator or brake pedals, referred to as the Firefly. Waymo also conducted limited public road rider-only testing using Chrysler Pacifica vehicles before the launch of the Waymo One ride-hailing service in December 2018. These operations did not involve any contact events but mileage from these operations are not included in this paper.

risk function described in McMurry et al. (2021) to compute the probability of MAIS2+ (p(MAIS2+)). The p(MAIS2+) was computed for hypothetical occupants in both the Waymo vehicle and other vehicles seated on both the left and right side of the vehicle. The highest p(MAIS2+) in either vehicle or in any seating configuration was taken as the overall event-level p(MAIS2+) reported in this study. Because the occupancy of the Waymo equipped vehicle and other party's vehicle can vary, making these occupancy assumptions may overestimate the actual extent of injuries based on the real vehicle occupants (e.g., when the ADS-equipped vehicle is unoccupied).

Comparison dataset selection and calculation of injury risk

Two nationally representative collision databases were analyzed to provide a point of reference with which to compare the Waymo contact events based on the types and severity of collisions. First, data was extracted from the Crash Investigation Sampling System (CISS) (National Highway Traffic Safety Administration, n.d.) case years 2017 through 2020. CISS is a nationally representative sample of police-reported collisions where at least one vehicle was towed from the scene. The collision investigated must involve a passenger vehicle. Collisions involving pedestrians and cyclists are generally not investigated. To provide a broader view of serious collisions, the Fatality Analysis Reporting System (FARS) years 2016 to 2020 was also analyzed. FARS provides a census of fatal motor vehicle collisions that occur in the U.S. It is common to pool several years of crash database data in an analysis to provide an estimate with less variance, especially when subsetting the data as done in this study.

To provide a relevant comparison, collisions from CISS and FARS were *filtered* to those that occurred on the types of roads that are within the ODD of Waymo's ride hailing operations: collisions were selected if they involved at least one passenger vehicle on non-highway roads with speed limit 45 mph and below, and with no adverse weather (e.g., snow or ice). We will refer to this sample as the CISS and FARS "ride hailing environment" collisions in this paper.

Furthermore, some collisions from the CISS database have a collision reconstruction performed that allow for injury risk to be estimated (approximately 40% of the CISS collisions within the ride hailing environment). An accident reconstructionist visits the scene of the collision and makes vehicle damage measurements to the involved vehicles. The investigators compute the same delta-V and PDOF described above using collision reconstruction techniques. Collisions that had a reconstruction performed were analyzed using the same injury risk function in McMurry et al. (2021) to estimate p(MAIS2+) to make a comparison with the severity of the Waymo contact events.

The resulting dataset after filtering for the ride hailing environment was composed of 2,345 collisions, with a weighted occurrence count of 1,731,218 from the CISS database and 38,527 fatal collisions from FARS.

Results

Twenty contact events were observed during the ADS's first one million miles of rider-only operations. Table 1 provides a description of the 20 contact events in chronological order. "Fleet" refers to the platform and location as described earlier: "PHX 4G" is Phoenix (East Valley) using the 4th generation platform, "PHX 5G" is Phoenix using the 5th generation platform, and "SF 5G" is San Francisco using the 5th Generation Platform. The conflict groups describe the maneuvers of the conflict partners, as described in Kusano et al. (2023). The four conflict groups that were observed in the Waymo rider-only contact events were "BACK" for backing, "F2R" for Front-to-Rear, "SV" for Single Vehicle, and "ODLI" for Opposite Direction Lateral Incursion. See Table A2.1 in the Appendix 2 for a listing and description of traffic conflict groups.

The first event that occurred had the highest severity of the 20 contact events: 4% as measured by p(MAIS2+). In this event, the Waymo vehicle was struck in the rear while slowing for a red light by a car driven by a teenage driver. Rearward facing video recorded by the ADS suggests the driver of the other vehicle was looking at a cell phone held near the steering wheel immediately prior to the collision.

Contribution

In evaluating the contact events shared in this paper, we encourage the reader to consider the nature of each event and the contribution each party made to that event. Contribution (or fault assessment) is a common concept in human-to-human collisions. It is widely used by the insurance industry to evaluate individual drivers' risk levels, by putting different weights on collisions with no contribution compared to collisions with significant contribution. It is also an important lens through which Waymo can prioritize engineering efforts to where improvements are most likely to be impactful for further enhancing the safety of autonomous driving technology. In reviewing events from the perspective of contribution, it is significant that every vehicle-to-vehicle contact event in the first one million miles involved one or more road rule violations or dangerous behaviors on the part of the operator of the other vehicle.

Table 1. Summary of Waymo Contact Events from one million Miles of Rider-Only (RO) Operations.

Event #	Year	Month	Fleet	Conflict Group	Conflict Partner	Reported Injury	p(MAIS2+)	Damage Description	Narrative Description
1	2020	9	PHX 4G	F2R	Passenger Vehicle	None reported	4.0%	Waymo AV: Rear bumper, liftgate, and radar Passenger car: Front bumper and hood, driver airbag deployed	Contact occurred between the rear of a Waymo AV and the front of a passenger car. The Waymo AV was coming to a gradual stop behind a queue of vehicles that were stopped at a red light when the passenger car behind the Waymo made contact with the rear of the Waymo AV. At the time of contact, the Waymo AV was traveling at approximately 5 mph and the passenger car was traveling at approximately 25 mph.
2	2021	4	PHX 4G	BACK	Passenger Vehicle	None reported	0.3%	Waymo AV: None Passenger car: None reported	Contact occurred between the right rear corner of a Waymo AV and the left rear corner of a passenger car. The Waymo AV was stopped in a parking lot to pick up a passenger when a passenger car backed out of a parking space and made contact with the Waymo AV. At the time of contact, the Waymo AV was stationary and the passenger car was traveling at approximately 1 mph.
3	2021	5	PHX 4G	SV	Object	None reported	0.0%	Waymo AV: None Construction pylon: None reported	Contact occurred between the front right of a Waymo AV and a construction pylon. The AV was traveling on a road with construction pylons between two lanes when the Waymo AV made contact with a construction pylon. At the time of contact, the Waymo AV was traveling at approximately 8 mph.
4	2021	5	PHX 4G	F2R	Passenger Vehicle	None reported	1.1%	Waymo AV: Dents/scratches to rear bumper and rear radar SUV: None reported	Contact occurred between the left rear of a Waymo AV and the front right corner of an SUV. While the Waymo AV was preparing to make a right turn on red onto a multilane road, an SUV made contact with the rear of the Waymo AV. At the time of contact, the Waymo AV was traveling approximate 1 mph and the SUV was traveling approximately 4 mph.
5	2021	9	PHX 4G	BACK	Passenger Vehicle	None reported	0.4%	Waymo AV: Dents/scratches on left rear door Passenger car: None reported	Contact occurred between the side of a Waymo AV and the rear of a passenger car. The Waymo AV was stopped in a parking lot to pick up passengers when a passenger car backed out of a parking space and made contact with the Waymo AV. At the time of contact, the Waymo AV was stationary and the passenger car was traveling at approximately 4 mph.
6	2021	10	PHX 4G	BACK	Passenger Vehicle	None reported	0.3%	Waymo AV: Dents/scratches on rear lift gate SUV: None reported	Contact occurred between the right rear corner of a Waymo AV and the left rear corner of an SUV. The Waymo AV was parked on a residential street when an SUV backed out of a driveway and made contact with the Waymo AV. At the time of contact, the Waymo AV was stationary and the SUV was traveling at approximately 2 mph.
7	2021	10	PHX 4G	F2R	Passenger Vehicle	None reported	2.0%	Waymo AV: Front bumper and front radar Passenger car: Rear bumper	Contact occurred between the front of a Waymo AV and the rear of a passenger car. The Waymo AV was on a multilane road approaching an intersection. As the traffic signal turned yellow, a passenger vehicle in an adjacent lane changed lanes in front of the Waymo AV and immediately braked. The Waymo AV exerted maximum brake force to avoid contact, but was unable to stop before making contact with the rear of the passenger vehicle, which had entered the intersection and was still braking. At the time of contact, the Waymo AV was traveling at approximately 21 mph and the passenger vehicle was traveling at approximately 10 mph.
8	2022	1	PHX 4G	BACK	Passenger Vehicle	None reported	1.2%	Waymo AV: None Passenger car: None reported	Contact occurred between the left front corner of a Waymo AV and the rear of a passenger car. The Waymo AV was stopped in a parking lot aisle to pick up passengers when a passenger car backed out of a parking space and made contact with the Waymo AV. At the time of contact, the Waymo AV was stationary and the passenger car was traveling approximately 1 mph.

Event #	Year	Month	Fleet	Conflict Group	Conflict Partner	Reported Injury	p(MAIS2+)	Damage Description	Narrative Description
9	2022	1	PHX 4G	BACK	Passenger Vehicle	None reported	1.2%	Waymo AV: None Passenger car: None reported	Contact occurred between the front of a Waymo AV and the right rear of a passenger car. The Waymo AV had stopped in a parking lot to yield to a passenger car that began to back out of parking space. The passenger car turned while backing out and the right rear of the passenger car made contact with the front of the Waymo AV. At the time of contact, the Waymo AV was stationary and the passenger car was traveling approximately 2 mph.
10	2022	4	PHX 4G	F2R	Passenger Vehicle	None reported	1.3%	Waymo AV: None SUV: None reported	Contact occurred between the rear of a Waymo AV and the front right corner of an SUV. While the Waymo AV was preparing to make a right turn from a red light onto a multilane road, an SUV made contact with the rear of the Waymo AV. At the time of contact, the Waymo AV was stationary and the SUV was traveling approximately 3 mph.
11	2022	4	SF 5G	BACK	Passenger Vehicle	None reported	0.3%	Waymo AV: None Passenger car: None reported	Contact occurred between the right side of a Waymo AV and the rear of a passenger car. The Waymo AV was stopped on a roadway yielding to a pedestrian that was walking a dog in a crosswalk when a passenger car that was parked perpendicular to the Waymo AV moved backward and made contact with the Waymo AV. At the time of contact, the Waymo AV was stationary and the passenger car was traveling less than 1 mph.
12	2022	6	PHX 4G	F2R	Passenger Vehicle	None reported	1.3%	Waymo AV: None Passenger car: None reported	Contact occurred between the rear of a Waymo AV and the front right corner of a passenger car. While the Waymo AV was beginning to make a right turn onto a multilane road from a stop sign, a passenger car made contact with the rear of the Waymo AV. At the time of contact, the Waymo AV was traveling at approximately 3 mph and the passenger car was traveling approximately 6 mph.
13	2022	8	PHX 4G	BACK	Passenger Vehicle	None reported	0.3%	Waymo AV: Scratches/dents on right rear corner Pickup: None reported	Contact occurred between the right rear corner of a Waymo AV and the rear of a pickup truck. The Waymo AV was stopped in a parking lot to drop off a passenger when a pickup truck backed out of a parking space and made contact with the Waymo AV. At the time of contact, the Waymo AV was stationary and the pickup truck was traveling approximately 1 mph.
14	2022	8	SF 5G	F2R	Passenger Vehicle	None reported	1.7%	Waymo AV: Scratches/dents on rear bumper SUV: Scratches/dents on front bumper and hood	Contact occurred between the rear of a Waymo AV and the front of an SUV. While merging onto a multilane road, the Waymo AV came to a stop to yield for traffic approaching from the left. An SUV approached from behind and made contact with the rear bumper of the Waymo AV. At the time of contact, the Waymo AV was stationary and the SUV was traveling at approximately 6 mph.
15	2022	10	PHX 5G	SV	Object	None reported	0.0%	Waymo AV: None Sign stand: None reported	Contact occurred between the side of a Waymo AV and a portable plastic sign stand. The Waymo AV was driving through an intersection and routing around a portable plastic sign stand lying in the roadway when the sign stand was blown by the wind and made contact with the side of the passing Waymo AV. At the time of contact, the Waymo AV was traveling approximately 11 mph and the sign stand was traveling approximately 3 mph.
16	2022	11	PHX 5G	ODLI	Heavy Vehicle	None reported	0.0%	Waymo AV: Left rear radar cover came off Garbage truck: none	Contact occurred between the left rear corner of a Waymo AV and the side of a garbage truck. The Waymo AV had pulled to the right on a narrow residential street, unable to proceed past an oncoming garbage truck and a parked vehicle. While attempting to pass the Waymo AV, the garbage truck made contact with the left rear corner of the Waymo AV. At the time of contact, the Waymo AV was stationary and the garbage truck was traveling less than 1 mph.

Event #	Year	Month	Fleet	Conflict Group	Conflict Partner	Reported Injury	p(MAIS2+)	Damage Description	Narrative Description
17	2022	12	PHX 5G	SV	Object	None reported	1.3%	Waymo AV: Scratches/dents on front and right side of vehicle Barrier arm: None reported	Contact occurred between a Waymo AV and a free-swinging parking lot barrier arm. While traveling in a parking lot, the front and right side of the Waymo AV made contact with a parking lot barrier arm. At the time of contact, the Waymo AV was traveling approximately 11 mph.
18	2022	12	PHX 5G	SV	Object	None reported	0.0%	Waymo AV: None Shopping cart: None reported	Contact occurred between a Waymo AV and a shopping cart. As the Waymo AV completed a right turn out of a parking lot, the right front corner of the Waymo AV made contact with a shopping cart that was in the roadway. At the time of contact, the Waymo AV was traveling approximately 13 mph.
19	2023	1	PHX 5G	BACK	Passenger Vehicle	None reported	0.4%	Waymo AV: Scratches on left rear radar cover bezel Tow truck: None reported	Contact occured between the left rear corner of the Waymo AV and the side of a tow truck. The Waymo AV was stopped in a parking lot when a tow truck maneuvering near the Waymo AV made contact with the the left rear corner of the Waymo AV. At the time of contact, the Waymo AV was stationary and the tow truck was traveling approximately 2 mph.
20	2023	1	PHX 5G	SV	Passenger Vehicle	None reported	0.0%	Waymo AV: Scratches on front bumper Minibike: Unknown if contact with Waymo resulted in additional damage	Contact occured between the front of a Waymo AV and a small motorcycle commonly known as a minibike. The Waymo AV was traveling in a multilane road with a group of motorcyclists traveling ahead in an adjacent lane. A minibike rider in the group lost control and fell off the minibike and the minibike tumbled into the Waymo AV's lane. The Waymo AV applied maximum brake force but made contact with the riderless minibike. At the time of contact, the Waymo AV was traveling approximately 8 mph and the minibike was traveling approximately 1 mph.

Patterns by conflict type

In this section, we describe some patterns observed in the contact events presented in Table 1 above. Numbers are based on Table 1 and are not mutually exclusive since each event may fit into more than one pattern.

Other Agent Backing

The single most common conflict type was another vehicle backing into the Waymo equipped vehicle (8 out 20). In all backing contacts, the Waymo vehicle was stationary at the time of impact. The Waymo vehicle was either stopped in the aisle of a parking lot to drop off or pick up passengers (5 of 8), or stopped on a road yielding to either the backing up vehicle or other road users (2 of 8), or parked on the side of the road (1 of 8).

Front to Rear

The next most common conflict type was front to rear (also sometimes called rear-ends) with 6 out the 20 contact events. In 5 out of those 6, the Waymo equipped vehicle was struck from behind. In these struck from behind contact events, the Waymo equipped vehicle was either stationary or moving slowly at the time of impact. In all struck cases, the Waymo equipped vehicle was yielding to a traffic light or other stopped vehicles. In 3 out of the 5 struck from behind cases, the Waymo Vehicle was stopped or moving forward preparing to turn right on a red traffic light. In 1 of the 5 struck-from-behind contacts the Waymo Vehicle was stopped yielding to traffic as it prepared to merge onto another roadway. In 1 of the 5 struck-from-behind contacts, the Waymo vehicle was coming to a gradual stop (no more than 0.2 g deceleration) to yield to traffic stopped at a red traffic light. There was 1 front to rear conflict where the Waymo equipped vehicle struck the rear of another vehicle after it changed lanes into the Waymo vehicle's lane and applied the brakes (see narrative description for event 7 in Table 1).

Contact with Objects in the Roadway

Five (5) of the 20 contact events were single vehicle incidents where the Waymo vehicle contacted an object that was in the roadway. In these contact events, the Waymo vehicle did not leave the roadway. At the time of impact, the Waymo equipped vehicle was traveling between 8 and 13 mph and the objects were either stationary or traveling at less than 3 mph. The objects contacted were all non-fixed objects with low mass: a construction pylon, shopping cart, swinging gate, plastic folding sign, and an unoccupied miniature motorcycle. This is in contrast with many police reported single vehicle collisions with objects that are fixed (e.g., poles or trees), which, as a result, pose a higher injury risk to the vehicle occupants compared to non-fixed, low mass objects.

Opposite Direction Lateral Incursion

The remaining event was an opposite direction lateral incursion conflict where a heavy vehicle (garbage truck) made contact at low speed to the side of a stopped Waymo equipped vehicle. This type of contact event occurred while negotiating a narrow passageway, and less similar to common opposite direction collisions in police-reported collisions that feature drivers drifting over the lane line while traveling at speed.

Comparison with nationally representative collision databases

As detailed in Table 1 above, the vast majority of the contact events are of minor severity, where nine of the 20 events resulted in no damage. To provide further context for the types and severities of the Waymo contact events, we computed the risk of potential occupants involved in each contact event receiving one or more MAIS2+ injuries as described in the methodology section. We then compared the Waymo contact events to a set of nationally representative police-reported collisions from CISS where at least one of the vehicles involved in the collision was towed from the scene of the collision. Further comparison of traffic conflict groups in the Waymo contact events, CISS, and FARS databases is included in Appendix 3. These types of “police accident report” databases are standard data sources used in traffic safety research. As mentioned before, police report databases such as CISS likely have underreporting, especially at lower severity collisions. Because of the advanced sensor suite, and operational capability of an ADS ride-hailing service like Waymo’s, our hypothesis is that there is a very high likelihood of detecting contact events. Thus, if all contact events are included in comparison with nationally representative crash data, they will have an overall lower severity in terms of risk of injury. Although there were no reported injuries in the Waymo contact events, presenting a risk of injury can allow for comparison of the severity of the Waymo contact events and police-reported collisions. To enable correct comparison with CISS inclusion criteria (police-reported collisions where at least one vehicle was towed from the scene) Waymo contact events were divided into CISS-comparable events and those that did not meet the criteria.

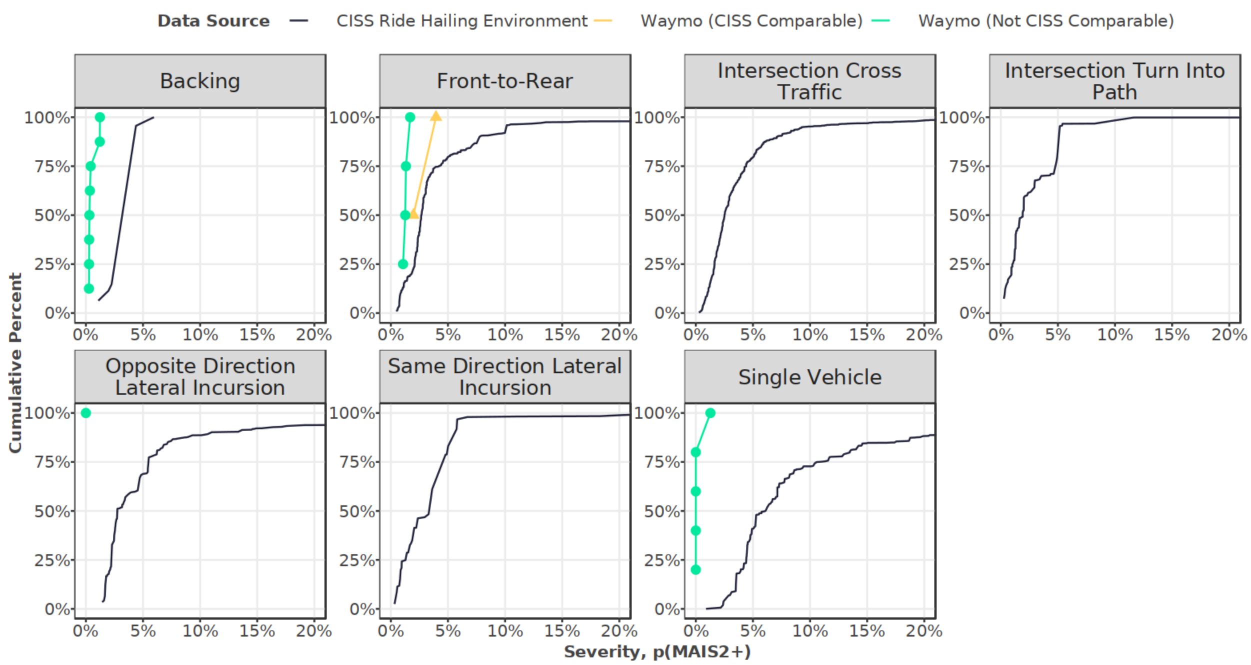


Figure 3. Cumulative Distribution of p(MAIS2+) in CISS Ride Hailing Environment and Waymo Contact Events by Conflict Group. The horizontal axis is the probability of a maximum Abbreviated Injury Scale of 2 or greater p(MAIS2+); and the vertical axis shows the cumulative percentage of each data source (CISS or Waymo events)

Figure 3 shows the cumulative distribution of p(MAIS2+) by the conflict groups for a) CISS "ride hailing environment" collisions, b) Waymo CISS-Comparable contact events, and c) minor contact events that did not meet the criteria for inclusion in CISS. The horizontal axis of each subplot shows the probability of a maximum Abbreviated Injury Scale of 2 or greater p(MAIS2+) and the vertical axis shows the cumulative percentage of each data source (CISS or Waymo events). The cumulative percentages sum to 100% for each data source, although the horizontal axis (p(MAIS2+)) is restricted to 20% to aid readability. When comparing the CISS and Waymo events, the cumulative distributions of the Waymo events are to the left of the CISS distribution, suggesting the Waymo events have a severity distribution with more probability density shifted towards lower p(MAIS2+). There were observed Waymo contact events in the Backing, Front-to-Rear, Opposite Direction Lateral Incursion, and Single Vehicle conflict groups. There was only 1 opposite direction lateral incursion Waymo contact event with p(MAIS2+) = 0%, which is represented as a single point at 100% cumulative percent.

The shapes of the cumulative distributions in Figure 3 confirm that most of the Waymo contact events do not meet the reporting requirements for CISS, described below. Events number 1 and 7, both Front-to-Rear contacts, met this requirement and are indicated as the "CISS-comparable" line in Figure 3. These two events have p(MAIS2+) values that fall approximately in the 15th percentile and 65th percentile of the CISS Front-to-Rear distribution. Although it is difficult to draw distributional differences from two data points, the events that are CISS comparable are not above 5% p(MAIS2+) for front-to-rear collisions. Further contextualizing these results, it can be seen in the comparison between FARS and CISS, in Appendix 3, that Front-to-Rear collisions represent a low proportion of the fatalities (6%) in comparison to CISS (19%). In the CISS reportable contact event #1, with the highest severity at 4% p(MAIS2+), the rearward facing video recording from the ADS suggested that the driver of the striking vehicle was looking down at a cell phone as the vehicle approached the slowing Waymo equipped vehicle.

The cumulative probability density of p(MAIS2+) for the “not CISS comparable” events is substantially lower than the police-reported data in backing, opposite direction lateral incursion, single vehicle collisions, and front-to-rear collisions with maximum p(MAIS2+) risk in Waymo “not CISS comparable” contact events lower than the minimum reconstructed CISS collisions.

Although weighted two-sample Kolmogorov-Smirnov (KS) tests show significance when comparing the “not CISS comparable” contact events to the CISS collisions (and also show significant differences when all 20 Waymo contact events are used), we choose not to report those results as they are not comparable datasets. CISS has a different inclusion criteria (the collision must involve at least one passenger vehicle, be police reported, and have at least one vehicle towed from the scene) than the Waymo events. Further, it is not statistically meaningful to perform a test with only two data points in the Waymo CISS-comparable dataset.

Discussion

The aim of this paper was (a) to examine all contact events experienced during the first one million miles of rider-only operations (with no human behind the wheel), and (b) explore what conclusions can be drawn from this observed real-world safety performance.

The contact events

During the first one million miles of rider-only operations of the Waymo One ride hailing service (i.e., no human behind the wheel), there were no reported injuries, only two CISS-comparable collisions (i.e., police reportable and a vehicle towed from the scene), and 18 contact events that were too minor to meet the criteria for inclusion in CISS. Most contact events resulted in little damage, with nine of the 20 events resulting in no damage. 55% of the contact events occurred with a stationary Waymo vehicle, and 40% were parking-related, with some overlaps between the two categories. Every vehicle-to-vehicle contact event involved one or more road rule violations or dangerous behaviors on the part of the human driver or the other actor.

One million rider-only miles is equivalent to over 80 years of driving for the average U.S. licensed driver (McGuckin & Fucci, 2018). No contact events involved VRUs or occurred in conflict groups that are often observed in fatal or police-reported collisions, such as intersection cross traffic, intersection turn into path, and crossing road (namely, pedestrians crossing the road and being struck by a motor vehicle). These four conflict groups make up 51% of all fatal collisions and 50% of CISS collisions in the ride hailing environment, but had no observed Waymo contact events. Single vehicle collision events make up 21% of fatal collisions.

Time of day

Waymo operates in rider-only mode 24 hours a day, seven days a week, and we have observed that the driving environment in our ODD varies significantly depending on the time of the day. Most notably, driving late at night (10 PM to 6 AM) results in fewer encounters with other road users than during other times of the day. For example, during these late night hours in San Francisco, the Waymo AV encounters an average of 85% fewer nearby cyclists, 83% fewer nearby motorcyclists, and 65% fewer nearby pedestrians than during other hours.

As expected given the less crowded driving environment observed late at night, Waymo experienced a lower frequency of contact events during these hours. During its first one million rider-only miles, 20% of miles were driven between 10 PM to 6 AM but only two of the twenty contact events (10%) occurred during these hours (events #4 and #5).

These results seem to contrast with data for human drivers, which some existing literature has concluded shows a relatively higher rate of injurious collisions at night versus during the day (Varghese & Shankar, 2007). There are a few possible explanations for this. One explanation is that despite a simpler driving environment, human drivers experience an increased rate of injurious collisions at night due to a higher prevalence at night of alcohol, fatigue, and speeding – risk factors that the Waymo Driver does not engage in at any hour. Another potential explanation is that human drivers also experience more contact events during the daytime than late at night, but that many daytime events are lower severity and are not reflected in higher-severity datasets like CISS and FARS. While this paper reports all contact events for the Waymo driver, when human drivers are involved in minor parking lot contacts and similar events (which may occur more frequently in congested daytime environments) they may go unreported.

Comparison with CISS “ride hailing environment” collisions

Waymo contact events were compared to standard data sources used in traffic safety research: a set of nationally representative police-reported collisions from the CISS database, where at least one of the vehicles involved in the collision was towed from the scene of the collision, and the FARS database for fatality collisions. Collisions from CISS and FARS were filtered to match those that occurred on the types of roads that are within the ODD of Waymo’s ride hailing operations.

The two CISS-comparable front-to-rear collisions are of low severity at 2% and 4% p(MAIS2+), in a conflict-type that is common in CISS (19%), but represents a lower involvement in fatalities (6%) in FARS. It is encouraging that when comparing the Waymo CISS-comparable collisions with the FARS and CISS databases, there were no contact events in conflict groups that represent 94% of the fatal collisions, and 81% of the CISS collisions (see Appendix 3). The distribution of severity (p(MAIS2+)) for the minor contact events is substantially lower than the police-reported data in backing, opposite direction lateral incursion, and single vehicle collisions. The severity for front-to-rear Waymo contact events, when combining both the CISS-comparable and minor contact events, was significantly lower than the police-reported data. This is an expected result because CISS is designed to under-report low-severity events, due to the tow away requirement for inclusion. Therefore, care must be taken to avoid making "apples to oranges" comparisons between nationally representative collision datasets and all Waymo contact events. Any comparison of CISS and the Waymo data sets would need to take into consideration the major differences in severity if all contact events were to be included in comparison.

Following the "mining the diamond" concept discussed in the introduction section, some lower severity events (like some of the contact events reported in this study) may not be relevant for the incidence of higher severity collisions. That is, lower severity events may not follow the original assumptions of relevance for higher severity levels of the Heinrich Triangle (Figure 2). For example, the most frequent collision mode reported from the Waymo contact events was another party backing into the Waymo vehicle in a parking lot or while the Waymo vehicle was parked on the roadside (8 out of 20 contact events). Given the relatively low speeds that are likely when vehicles are starting from standstill, a high severity outcome is unlikely. In contrast, other collision modes, such as intersection cross traffic conflicts like other vehicles crossing the path or turning in front of the ADS-equipped vehicle, have a much higher potential for injurious outcomes than these backing, parking-related conflicts. The data supports a "mining the diamond" approach to analyzing non-injury contact events, whereby focus should be placed on those events that have the greatest potential to result in injury. Correlation between surrogate metrics and serious outcomes might be achievable if only those surrogate events that could potentially result in a serious outcome are considered. Metrics like the maximum injury potential (Kusano & Victor, 2022), which estimates the worst-case injury outcome in a traffic conflict if both actors were to continue at the same speed and acceleration without avoidance maneuvers, is a potential quantitative metric to identify high injury potential events for use in surrogate metrics. Further understanding here will broadly benefit traffic safety.

Curating an ODD-specific dataset to create an appropriate benchmark to compare the ADS performance is challenging for many reasons, such as differences between humans and an ADS in behavior and route selection, driver and traffic characteristics, etc. (see e.g. Lindman et al., 2017), and the ODD dilemma - the tradeoff between using large, robust datasets (e.g., nationally representative databases) and small, ODD-specific datasets that have fewer observations and do not contain crashes representative of high severity (Scanlon, et al., 2021). The aforementioned reporting differences in police report or insurance databases also present a challenge to establish a benchmark. While NDS datasets have video and other sensor data that makes a comparison to low severity collisions possible, even the largest NDS lack a large number of serious injury or fatality collisions to establish a benchmark. Police report data, like the CISS database, have a number of serious injury and fatal collisions, but suffer from underreporting of low severity collisions (Blincoe et al., 2015). In addition to finding data with comparable severity, estimating the driving mileage or time (the denominator to a benchmark rate) is equally as challenging. Various data sources from local, state, and federal agencies estimate Vehicle Miles Traveled (VMT), like the Highway Performance Monitoring System (HPMS) administered by the Federal Highways Administration. The VMT data is most commonly estimated from temporary field surveys, continuous traffic volume monitoring, floating car surveys, or traffic flow simulation. These VMT studies are difficult to perform and are usually constrained by budget, for example to conduct the field surveys, which may lead to low precision estimates especially for the low functional class roads (e.g., local roads) that an ADS ridehailing vehicle frequently drives on.

Therefore, calculating either injury reduction or a reduction in the rate of collisions using police report or insurance claims data to compare with the minor contact events in this paper is inappropriate as they do not have equal reporting characteristics (particularly regarding severity levels) and may also not be collected from a comparable ODD. The results in this study show that the vast majority of the Waymo contact events are of such low severity that they do not result in injury or even property damage. By estimating the injury risk (p(MAIS2+)) of the Waymo contact events, we enabled a comparison with reference collision databases to contextualize the severity of these events. The current ADS crash reporting requirements make it difficult to assess the severity of collisions, one of two key indicators of risk reduction (occurrence of harm and severity of harm). A standardized method to estimate injury risk that is generalizable to all collision modes and partners is desirable, but is not a trivial task, and would require stakeholder alignment. For example, estimating the mass of the other contacted object from sensor data can be challenging. These same challenges exist for estimating injury risk from benchmark data sources, such as NDS, where the sensors may have blind spots that make it difficult to determine if contact was made. Likewise, to make valid comparisons between ADS systems, it's important to account for variables in driving conditions that may impact safety, such as road types, traffic and pedestrian density, weather, and hours of operations.

Limitations and future work

Due to the rarity of high-severity collision scenario occurrences, it is difficult to rely on actual real-world, on-road occurrences of collisions before large-scale deployment. Instead, predictive estimates and safety-by-design processes are needed for safety assurance before readiness determination (as described in Webb, et al., 2020). Prospective methods (e.g. a predictive model) of estimating collision rates can be useful as leading measures of ADS safety if they can be validated within the applicability and the trustworthiness of the underlying data. Today Waymo can rely on over ten-years worth of data collected on the road, along with data from a state of the art simulation infrastructure, and now one million miles of real-world rider-only data. At the time the readiness determination is made, actual and simulated performance are combined to measure how safety performance indicators, upon which acceptance criteria are predicated, compare to specified targets and approval guidelines, to establish confidence in the ability to safely scale to the next level of deployment (see Webb, et al., 2020). The elaboration and validation of predictive models used in Waymo's predicted collision rates methodologies is a continuing topic outside the scope of the current paper.

The purpose of comparing the calculated p(MAIS2+) risk for the Waymo contact events and the CISS ride hailing environment was to be able to demonstrate that the Waymo contact events were of lower severity than the CISS events. The risk curves used to predict p(MAIS2+) are statistical models that were trained on police reported data (McMurry et al., 2021). Because some contact events go unreported to police (Blincoe et al., 2015), this data censoring creates the opportunity for overestimation of p(MAIS2+) risk on the very low injury severity end. Injury risk models like the ones used in this study have been traditionally used to compare risk across different populations (e.g., to predict the total number of injuries before and after the introduction of a safety system) where the predictions are a difference or reductions in injuries rather than the absolute p(MAIS2+) values displayed in this paper. As stated previously, many factors such as seat belt use, occupant characteristics, and vehicle characteristics will determine the injury outcomes in two collision populations. The p(MAIS2+) used the default predictors based on population averages described in McMurry et al. (2021) as a means of making a comparison between the two datasets. In order to predict the expected number of injured occupants these other factors (e.g., vehicle occupancy, seat belt use, occupant age) would need to be considered, which was not the purpose of the analysis done in this paper.

Continuous confidence growth

Waymo takes three interconnected perspectives on safety in its Safety Determination Lifecycle: safety as an emergent development property, safety as an acceptable prediction and/or observation, and safety as continuous confidence growth. As the deployment scale increases and the available data grows in volume, the statistical confidence of our determination of AUR improves. In order for such confidence to grow, it is fundamental to understand the difference in the *validity of predictions* versus the *credibility of observed occurrences* upon which those predictions can be validated. The ability to validate predictions increases as exposure increases (i.e., with accumulation of mileage). High credibility in rider-only-observed occurrences will be achieved with sufficient exposure (mileage and/or hours of operation). Thus, the notion of a confidence build-up combined with a rigorous credibility assessment process are central to confidence in statements based on safety outcomes. In other words, confidence in conclusions about the safety of the Waymo Driver increases when both validation of predictions and credibility of available data increase over time[7]. Valid assessment of a-posteriori, retrospective performance of the ADV, is most efficiently achieved with a comparative assessment to human driving performance.

Confidence in conclusions about the safety of the Waymo Driver increases when both (statistical) credibility of available data and validity of predictions increase over time. To tackle the credibility irony – wherein lower occurrence rates (better performance) leads to a larger required mileage to reach 100% credibility – it is necessary to (a) apply the same requirements

[7] The holistic determination of absence of unreasonable risk also relies on the rigorous application of engineering practices that are at the backbone of the "safety as an emergent development property". perspective. Waymo's collection of Acceptance Criteria span a number of performance indicators beyond the contact events presented in this paper

to ADS and human comparison, only counting events that are likely included in human reported crash data, and (b) requiring the same mileage target to reach the full credibility standard.

Conclusions

What conclusions can be drawn from the frequency and low severity of these contact events about the real-world safety performance of the Waymo Driver? During the first one million miles of rider-only operations of the Waymo One ride hailing service (i.e., no human behind the wheel), there were no reported injuries, and only two CISS-comparable collisions (i.e., police reportable and a vehicle towed from the scene), both of which were Front-to-Rear. Although Front-to-Rear collisions represent a frequently occurring conflict type in CISS (19%), they have a lower proportion of fatalities (6%) in FARS. In the more severe of the two, the Waymo was struck in the rear by a car driven by a driver looking at their cell phone while slowing for a red light. When comparing the Waymo CISS-comparable collisions with the FARS and CISS databases, there were no collisions in conflict groups that represent 81% of the CISS collisions and 94% of the fatal collisions. This report also includes 18 contact events that were too minor to meet the criteria for inclusion in CISS, nine of which resulted in no damage.

The results of this paper are consistent with the assertion that the Waymo Driver will reduce the frequency of severe collisions by (i) mitigating the potential for injury by driving safely, responsibly, and defensively to avoid entering into a conflict situation in the first place and (ii) taking an appropriate avoidance maneuver if a conflict develops. The current study, in combination with previous reported data (Schwall, et al., 2020) and past studies performed by Waymo on fatalities reconstruction (Scanlon et al., 2021, 2022) and on collision avoidance testing (Kusano et al., 2022), support the assertion that the Waymo driver is successful at reducing injuries and fatalities today, particularly through eliminating or reducing the frequency of fatal and serious collisions. The results of the Waymo contact events reported in this paper contribute, alongside many other methodologies, performance indicators, and state of the art development practices, to Waymo's determination of absence of unreasonable risk. Further quantitative validation from a-postori on-road outcomes will provide continuous growth of confidence in our system performance. To reach full confidence, that is, full statistical credibility, application of the same requirements to ADS and human comparison by only counting events that are likely included in a human reported crash data benchmark, is most efficient and therefore necessary.

References

Abebe, H.G., Hansson, S.O., Edvardsson Björnberg, K. (2022). Arguments Against Vision Zero: A Literature Review. In: Edvardsson Björnberg, K., Belin, MÅ., Hansson, S.O., Tingvall, C. (eds) The Vision Zero Handbook. Springer, Cham. https://doi.org/10.1007/978-3-030-23176-7_3-1

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., & Sayed, T. (2021). A systematic mapping review of surrogate safety assessment using traffic conflict techniques. *Accident Analysis & Prevention*, *153*, 106016.

Association for the Advancement of Automotive Medicine (2016). *The Abbreviated Injury Scale 2015 Revision*. Des Plaines, IL.

Bailey, A. L. (1945). A generalized theory of credibility. In *Proceedings of the Casualty Actuarial Society* (Vol. 32, No. 13, pp. 13-20).

Cai, Q., Abdel-Aty, M., Lee, J., & Eluru, N. (2017). Comparative analysis of zonal systems for macro-level crash modeling. Journal of safety research, 61, 157-166.

California Code of Regulations. Title 13, Article 3.7, Section 227.48. (2014). Testing of Autonomous Vehicles, Reporting Collisions. Accessed from https://www.dmv.ca.gov/portal/file/adopted-regulatory-text-pdf/

Carlson, A. (August 5, 2021). [In Re: First Amended Standing General Order 2021-01]. Retrieved from https://www.nhtsa.gov/sites/nhtsa.gov/files/2021-08/First_Amended_SGO_2021_01_Final.pdf

Chang, A., Saunier, N., & Laureshyn, A. (2017). *Proactive methods for road safety analysis* (No. WP-0005). SAE Technical Paper.

Blanco, M., Atwood, J., Russell, S. M., Trimble, T. E., McClafferty, J. A., & Perez, M. A. (2016). Automated vehicle crash rate comparison using naturalistic data. Virginia Tech Transportation Institute.

Blincoe, L., Miller, T. R., Zaloshnja, E., & Lawrence, B. A. (2015). *The economic and societal impact of motor vehicle crashes, 2010 (Revised).* National Highway Traffic Safety Administration, Report No. DOT HS 812 013.

European Union (August 5, 2022). *Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles.* Accessed from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1426

Favarò, F. (2021). Exploring the Relationship Between" Positive Risk Balance" and" Absence of Unreasonable Risk". arXiv preprint arXiv:2110.10566. https://arxiv.org/ftp/arxiv/papers/2110/2110.10566.pdf

Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018). *Measuring automated vehicle safety: Forging a framework*. RAND Corporation, Document Number RR-2662. https://www.rand.org/pubs/research_reports/RR2662.html

Gennarelli, T. A., & Wodzin, E. (2015). The abbreviated injury scale 2015.

Grembek, O., Vayalamkuzhi, P., Oum, S. (2023). Highway Crashes in California During the COVID-19 Pandemic: Insights and Considerations. In: Loukaitou-Sideris, A., Bayen, A.M., Circella, G., Jayakrishnan, R. (eds) Pandemic in the Metropolis. Springer Tracts on Transportation and Traffic, vol 20. Springer, Cham. https://doi.org/10.1007/978-3-031-00148-2_11

Heinrich, H.W. Industrial Accident Prevention – A Scientific Approach, 3rd Edition, 1950.

International Organization for Standardization (2018). Road Vehicles - functional safety. ISO 26262:2018.
International Organization for Standardization (2022). Road Vehicles - safety of the intended functionality. ISO 21448:2022.
Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?. *Transportation Research Part A: Policy and Practice*, *94*, 182-193.
Korner, J. (1989). A method for evaluating occupant protection by correlating accident data with laboratory test data. *SAE transactions*, 870-884.
Kullgren, A. (2008, September). Dose-response models and EDR data for assessment of injury risk and effectiveness of safety systems. In *Proc of Int. IRCOBI Conf., Bern, Switzerland* (pp. 3-14).
Kullgren, A., Lie, A., & Tingvall, C. (2010). Comparison between Euro NCAP test results and real-world crash data. *Traffic injury prevention*, *11*(6), 587-593.
Kusano, K., & Victor, T. (2022). Methodology for determining maximum injury potential for automated driving system evaluation. *Traffic injury prevention*, 1-4. https://doi.org/10.1080/15389588.2022.2125231
Kusano, K., Scanlon, J., Brännström, M., Engström, J., Victor, T. (2023). Framework for a Conflict Typology Including Contributing Factors for use in ADS Safety Evaluation. *In Proceedings of the Enhanced Safety of Vehicles Conference*. Yokohama, Japan. In Press.
Lindman, M., Ödblom, A., Bergvall, E., Eidehall, A., Svanberg, B., & Lukaszewicz, T. (2010). Benefit estimation model for pedestrian auto brake functionality.
Lindman, M., Isaksson-Hellman, I., & Strandroth, J. (2017, September). Basic numbers needed to understand the traffic safety effect of automated cars. In *IRCOBI Conference* (pp. 1-12).
Mahler, H., & Dean, C. (2001). Credibility. In *Foundations of Casualty Actuarial Science* (Chapter 8). Casualty Actuarial Society.
Manuele, F. A. (2011, October) Dislodging Two Myths From the Practice of Safety. *Professional Safety*.
Meyer, M.W. COVID Lockdowns, Social Distancing, and Fatal Car Crashes: More Deaths on Hobbesian Highways?. Camb J Evid Based Polic 4, 238–259 (2020). https://doi.org/10.1007/s41887-020-00059-8
McMurry, T. L., Cormier, J. M., Daniel, T., Scanlon, J. M., & Crandall, J. R. (2021). An omni-directional model of injury risk in planar crashes with application for autonomous vehicles. *Traffic injury prevention*, *22*(sup1), S122-S127.
Motor Vehicle Crash Data Collection. Accessed January 12, 2023 from https://www.nhtsa.gov/crash-data-systems/crash-investigation-sampling-system.
National Highway Safety Administration (n.d.). Crash Investigation Sampling System
Norberg, R. (2004). Credibility theory. *Encyclopedia of Actuarial Science*, *1*, 398-406.
Redelmeier, D. A. (2014). The fallacy of interpreting deaths and driving distances. *Medical Decision Making*, *34*(8), 940-943.
SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. (2021). https://doi.org/10.4271/J3016_202104

Scanlon, J. M., Kusano, K. D., Daniel, T., Alderson, C., Ogle, A., & Victor, T. (2021). Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. *Accident Analysis & Prevention*, *163*, 106454.

Scanlon, J. M., Kusano, K. D., Engstrom, J., Victor, T. (2022). Collision Avoidance Effectiveness of an Automated Driving System Using a Human Driver Behavior Reference Model in Reconstructed Fatal Collisions.

Schwall, M., Daniel, T., Victor, T., Favaró, F., & Hohnhold, H. (2020). Waymo public road safety performance data. *arXiv preprint arXiv:2011.00038*. https://arxiv.org/abs/2011.00038

Smith, R. M., & Jones, M. L. (2013, March). A Hurt-Based Approach to Safety. In *SPE Americas E&P Health, Safety, Security and Environmental Conference*. OnePetro.

Varghese, C. & Shankar, U. (May 2007) NHTSA Traffic Safety Facts, Passenger Vehicle Occupant Fatalities By Day and Night. National Highway Traffic Safety Administration, Report Number DOT HS 810 637.

Webb, N., Smith, D., Ludwick, C., Victor, T., Hommes, Q., Favaro, F., ... & Daniel, T. (2020). Waymo's safety methodologies and safety readiness determinations. *arXiv preprint arXiv:2011.00054*.

Yorio, P. L., & Moore, S. M. (2018). Examining factors that influence the existence of Heinrich's safety triangle using site-specific H&S data from more than 25,000 establishments. *Risk analysis*, *38*(4), 839-852.

Young, R. (2021). Critical Analysis of Prototype Autonomous Vehicle Crash Rates: Six Scientific Studies from 2015–2018. SAE International.

Appendix 1: Credibility

The notion of credibility plays a central role in the insurance sector, where it sits within common actuarial practices for rate making. Credibility theory is an established branch that goes back to seminal studies in the field of insurance mathematics (Norberg, 2004) (Bailey, 1945). Within this sector, credibility is more closely tied to the notion of confidence in predicting future outcomes, as informed by the applicability, believability, and the trustworthiness of the underlying data available for such an estimate. Traditional insurance applications often use occurrence counts such as the number of events or exposure count such as the number of miles driven as the basis for deriving the credibility of an observed rate of occurrence. For example, to estimate a fleet's expected crash count for the next year, average annual crash counts from historical years $E[N]$ is often used as the "estimator". In its most simplistic formulation, the Classical Credibility Framework, (Mahler & Dean, 2001) we can measure how accurate this estimator $E[N]$ is, by making statements like "the probability of the observed data N being within 5% of $E[N]$ is at least 90%", or $Pr\{\ (100\% - 5\%)E[N] \leq N \leq (100\% + 5\%)E[N]\ \} \geq 90\%$ in mathematical form. The minimum $E[N]$ needed for this to hold true, is then called the full credibility standard n_f. In practice, the confidence level p (90%) and the margin of error k (5%) can be judgmentally selected, and the full credibility standard n_f will change accordingly. Under the Classical Credibility Framework stated above, a commonly-used full credibility standard for occurrence count is n_f = 1082, based on p = 90%, k = 5% and assuming the occurrence rate follows a Poisson distribution and the normal approximation to the Poisson (Mahler & Dean, 2001). If we want to translate 1082 crashes into corresponding mileage, assuming a crash frequency c representing the expected crash count per mile, we would need $m = 1082/c$ miles to observe that many crashes.

Therefore, using historical observations to estimate crash frequency in this manner presents a challenge to any emerging technology that aims to reduce crash frequency by improving safety, including autonomous driving. The more safety benefits a technology brings, the lower the crash occurrence rate (smaller c) becomes, thus more miles ($m = 1082/c$) are needed to reach the full credibility standard. This problem is further magnified when it comes to rare events like high severity crashes; it is very difficult to derive a trustworthy high severity crash occurrence by merely relying on empirical data. However, benchmark-comparative methods to an agreed upon threshold, such as human driving performance, provide a framework that can lead to statistically significant conclusions on safety performance by drawing insights from differences, without the need to have a 100% credible crash occurrence estimate.

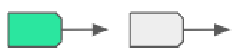
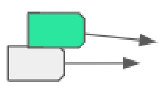
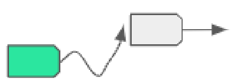
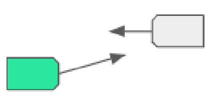
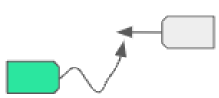
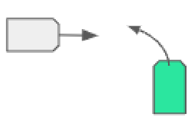
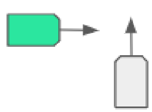
As discussed in Schwall et al. (2020), when comparing ADS driving performance to a selected benchmark, the mileage needed to reveal statistically significant differences also depends on the magnitude of the differences in the actual rates being compared. For a given metric, the larger the difference in performance, the fewer miles that are required to establish statistical confidence in a hypothesis of non-inferiority or superiority. Hence, it is also reasonable to state that ADS empirical crash occurrence rate deserves at least the same level of "believability" as the selected benchmark, especially when there is large performance difference. As an illustrative example, if we have concluded through credibility theory that a human driving set of exposures is fully credible at 100M miles, it should stand to reason that an ADS should achieve the same “believability” of its outcomes for the same set of miles driven, even if there are few to no crashes (a detectable difference can come earlier).

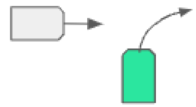
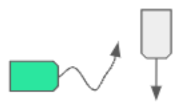
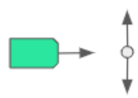
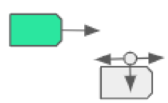
In sum, comparative methods can empower conclusions about the ADS safety performance much sooner, without the burden of having to derive complete collision frequency distributions from mileage accumulation. That is, through comparison with a robust human baseline, detection of safer performance can be done with fewer collision observations, and therefore lower mileage. This of course introduces the requirement for a robust benchmark.

Appendix 2: Conflict Groups

Table A2.1 summarizes the conflict groups introduced in Kusano et al. (2023). The conflict groups are one of the layers of a conflict typology that also describes the conflict partners, role (initiator or responder), and perspectives of each actor involved in a conflict.

Table A2.1. Conflict Groups and Short Descriptions (reproduced from Kusano et al., 2023, in press). A 🚙 symbol means the conflict group is relevant for conflicts between roadway actors. A 🚶 symbol means the conflict group is relevant for conflicts between a roadway actor and a non-roadway actor.

Conflict Group	Picture	Description
Single Vehicle (SV) 🚙 🚶		Includes all actions (or lack thereof) where the ego vehicle is traveling in a trafficway but then experiences an in-trafficway interaction without a conflict partner (e.g., a rollover event) or an off-trafficway interaction (e.g., a road departure).
Front-to-Rear (F2R) 🚙		Involves one road user interacting with another road user in the same direction and same travel lane.
Same-Direction Lateral Incursion (SDLI) 🚙		Occurs when two roadway actors are traveling in the same trafficway but in initially different travel lanes at the time of the initial interaction due to lateral incursion by some actor.
Same-Direction Prior Circumstances (SDPC) 🚙		Involves two roadway actors operating on the same trafficway in the same direction when one road user performs a lateral evasive action, experiences loss of control, or is involved in a prior collision that results in an interaction with the other road user.
Opposite Direction Lateral Incursion (ODLI) 🚙		Occurs when a non-turning actor operating in the trafficway's intended travel direction interacts with another actor that is operating opposite of the travel direction in the same trafficway.
Opposite Direction Prior Circumstances (ODPC) 🚙		Involves two roadway actors traveling in opposite direction trafficways in their respective trafficway's direction of travel when one road user performs a lateral evasive action, experiences loss of control, or is involved in a prior collision that results in an interaction with the other road user.
Turn into Path Opposite Direction (TIPOD) 🚙		Occurs as a result of one actor changing vehicle-operated trafficways via a turning maneuver and interacting with another actor, where one of these actors is operating in the opposite direction of the trafficway's direction of travel.
Intersection Cross Traffic (ICT) 🚙		Involves interactions that occur as a result of both actors changing or crossing over trafficways, and where the two actors cross paths with one another.

Conflict Group	Picture	Description
Intersection Turn Into Path (ITIP) 🚙		Involves interactions that occur as a result of one of the actors moving on to a trafficway via a turning maneuver into the path of another actor that is operating in the trafficway being turned on to.
Perpendicular Direction Prior Circumstances (PDPC) 🚙		Involves two roadway actors operating on crossing roadways that interact with one another following some lateral evasive action, prior loss of control, or prior collision.
Crossing Road 🚶		Involves interactions between an actor moving along a trafficway and another actor crossing that trafficway (while not traveling along or onto another trafficway).
Forward 🚶		Involves vehicle actors moving in the forward direction and interacting with a non-road user conflict partner in the trafficway that is not attempting to cross the road.
Interacting in Trafficway 🚶		Occurs when a forward moving ego is on a trafficway and interacts with an agent that is in the trafficway and moving around, entering, exiting, or interacting with an immediately adjacent vehicle or object.
Backing 🚙 🚶		Includes all interactions where at least one road user is moving in reverse.
Miscellaneous Circumstances		Events that do not fit into the aforementioned conflict groups, and are intended to cover all abnormal circumstance interactions that pose some collision risk.
Other/Unknown		All remaining events that do not fit into a conflict group, but that may need future considerations and those cases that have insufficient information to adequately determine the conflict group.

Appendix 3: Comparison with FARS and CISS

To provide additional context for comparison with serious collisions, Table A3.1 shows the proportion of events by conflict group for the CISS police-reported collisions, FARS fatal collisions and the Waymo CISS-comparable events. Because CISS does not currently include pedestrian or cyclist collisions, the proportion of conflict groups reported in CISS does not represent the overall crash population. To provide this broader view, we present the FARS data, which is a census of fatal collisions in the US. The percentages within each column sum to 100%, showing the relative frequency of occurrence of each conflict group in each data source. When comparing the Waymo CISS-comparable collisions with the FARS and CISS databases, there were no collisions in conflict groups that represent 94% of the fatal collisions, and 81% of the CISS collisions.

Table A3.1. Proportion of Police Reported Tow Away Collision (CISS) in the Ride Hailing Environment, Fatal Collisions (FARS) in the Ride Hailing Environment, and Waymo rider-only CISS-comparable events.

Conflict Group	% Police Reported + Tow Away (CISS)	% Fatal Collisions (FARS)	% Waymo RO CISS-comparable Events (Count)*
Intersection Cross Traffic (ICT)	43%	27%	0%
Single Vehicle (SV)	25%	21%	0%
Front to Rear (F2R)	19%	6%	100(2)%
Intersection Turn into Path (ITIP)	4%	1%	0%
Same Direction Lateral Incursion (SDLI)	3%	1%	0%
Opposite Direction Lateral Incursion (ODLI)	2%	8%	0%
Backing (BACK)	0.3%	0.2%	0%
Pedestrian Crossing Road (CR)	0%	22%	0%
All Others	4%	14%	0%
Total	100%	100%	100% (2)

* The Waymo contact events that did not meet the CISS-comparable event threshold were Backing (n=8), Front to Rear (n=4), Single Vehicle (n=5), and Opposite Direction (n=1). Because these events are not of the same reporting standard, they should not be directly compared to the CISS and FARS collisions in the table.