Fractional Collisions: A Framework for Risk Estimation of Counterfactual Conflicts using Autonomous Driving Behavior Simulations

Sreeja Roy-Singh, Sarvesh Kolekar, Daniel P. Bonny, Kyle Foss

Nuro AI, 1290 Terra Bella Avenue, Mountain View, CA 94036

[sreeja, skolekar, dbonny, kfoss] @nuro.ai

Abstract

We present a methodology for estimating collision risk from counterfactual simulated scenarios built on sensor data from automated driving systems (ADS) or naturalistic driving databases. Two-agent conflicts are assessed by detecting and classifying conflict type, identifying the agents' roles (initiator or responder), identifying the point of reaction of the responder, and modeling their human behavioral expectations as probabilistic counterfactual trajectories. The states are used to compute velocity differentials at collision, which when combined with crash models, estimates severity of loss in terms of probabilistic injury or property damage henceforth called fractional collisions. The probabilistic models may also be extended to include other uncertainties associated with the simulation, features, and agents. We verify the effectiveness of the methodology in a synthetic simulation environment using reconstructed trajectories from 300+ collision and near-collision scenes sourced from VTTI SHRP2 database and Nexar dashboard camera data. Our methodology predicted fractional collisions within 1% of ground truth collisions. We evaluate agent-initiated collision risk of an arbitrary ADS software release by replacing the naturalistic responder in these synthetic reconstructions with an ADS simulator and comparing the outcome to human-response outcomes. Our ADS showed reduced the naturalistic collisions by 4x and fractional collision risk by 62%. The framework's utility is also demonstrated on 250k miles of proprietary, open-loop sensor data collected on ADS test vehicles, re-simulated with an arbitrary ADS software release. The ADS initiated conflicts that caused 0.4 injury-causing and 1.7 property-damaging fractional collisions, and the ADS improved collision risk in 96% of the agent-initiated conflicts.

1. Introduction

The proliferation of automated driving systems (ADS) on shared public roads is expected to be supported by software releases that are evaluated to ensure safe and explainable behavior [19] [17]. The notion of sufficiently safe often measured by the "Absence of Unreasonable Risk" (AUR) in the United States and European Union and using "As Low As Reasonably Practicable" (ALARP) for the United Kingdom. The predictability or safety of ADS are also often compared to human-like driving to demonstrate a Positive Risk Balance (PRB), i.e. the proposition that an ADS should be at least as safe as (and ideally safer than) a human driver [23]. Since sole reliance on onroad testing to evaluate ADS is prohibitively expensive [20], simulation-based tests using sensor data collected onroad subsequently re-simulated with an ADS software test release have become popular. Simulation tests are not only more scalable and safe, they can also target testing of failure modes known to plague ADS more than humans [10]. The most widely known application of counterfactual simulations in ADS evaluation is to assess the counterfactual outcome of onroad disengagements by safety drivers, whereby the post-disengagement ADS behavior is simulated as if the safety driver not taken over from the ADS [37][42].

This paper presents a framework that provides high-fidelity estimates of collision risk in counterfactual simulations and supports both open- and closed-loop conflict data. As described in Figure 1, it is applicable to any pair of agent trajectories (B and C) in a simulation environment (A) via open-loop predictions after the point of conflict initiation. While we demonstrate the framework (D) in terms of responder behavioral uncertainty only, the underlying probabilistic models may be extended to include other uncertainties associated with the ADS simulator (e.g. controller errors in C), agents (e.g. aggressiveness in B), environment (e.g. latency in A). We also evaluate the framework and the underlying models using a novel method that uses naturalistic aggregate collision risk as ground truth. This method can be used by developers to evaluate any simulator or agent models used in the framework beyond the models described here. Finally, we show the framework's impact to ADS evaluation over large volumes of scenarios collected on our ADS test fleet or naturalistic databases.
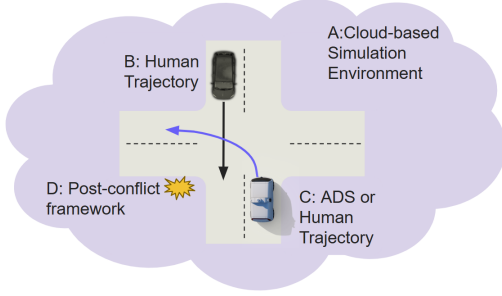
1

Figure 1. Simulation overview in a 2-body conflict. Agent C initiates a conflict and agent B is expected to respond to avoid a collision. For evaluating human-human performance, both trajectories are reconstructed from naturalistic databases, and for human-ADS performance, C's trajectories come from the ADS simulator and B's trajectories from either logged sensor data by an ADS test fleet or naturalistic reconstructions. This paper focuses on a framework (D) to model B's responsive behavior *after* conflict onset in any sim environment (A) and any pre-conflict trajectory (B,C)

2. Motivation and State of Art

Accurate estimation of ADS-initiated collision risk requires modeling of agents that react to the ADS. Closed-loop sensor simulation on controllable simulators with reactive agents allows for dynamic evaluation of the ADS with its environment and actors [46][15], although realism of the world has been questioned [10]. Open-loop simulation replays a non-adaptive environment around the simulated ADS (Fig1-C), at significant cost savings and benefits of real-world-based conflict triggers, but requires us to explicitly model the behavior of the reacting agent (Fig1-B). The ADS can be simulated open-loop in generative synthetic environments with line following agents or using logged onroad data, where the simulated ADS can perceive the agents/objects around it and respond to them. However, the agents whose trajectories are replayed *do not or cannot* respond to the ADS's behavior, which may change from the logged version with every software release. To address this *sim2real* gap, we use counterfactual simulations to model the possible behavioral reactions that a human responder may have if faced with that realistic conflict. Post-conflict counterfactuals are also relevant for hybrid simulators where some components are closed-loop (e.g. pre-conflict nominal agent behavior, sensor sim) but collision evasiveness is insufficiently modeled.

Accurate PRB-benchmarking of agent-initiated collision risk requires modeling of agents that react to the initiator. Open-loop counterfactuals provide dense, realistic signal for agent-initiated conflicts. Such conflicts are rare on public roads because safety drivers in ADS test vehicles drive defensively, therefore simulations are the preferred test mode for coverage (or structured tests on a closed course). Pose divergence between the ADS and the logged

vehicle after safety disengages motivates counterfactual situations. Using Figure1 for a different use case, let's assume ADS test vehicle (C) had right of way in the onroad test, but the safety driver disengaged the ADS and yielded to B, causing B to proceed first. When the counterfactual is simulated, if the ADS (C) in the virtual environment asserts its right of way and the (non-reactive) agent is found to violate it, the scenario becomes a useful example of an agent-initiated conflict. Collision avoidance testing using accident reconstructions have benchmarked collision risk, e.g, for Waymo Driver [36] and Baidu Apollo ADS [47]. However, they typically do not employ human behavior models because one agent's trajectory is played back exactly in the simulator while the other agent is replaced by the ADS. To address this *generic benchmarking* gap, we develop probabilistic counterfactuals for human behavior to quantify their collision avoidance in place of the simulated ADS.

Behavior modeling of reacting agents should be explainable and their outcomes comparable to the real world. Formal verification methods for scoring the risk of an ADS dynamic driving task (DDT) with respect to static obstacles and dynamic traffic agents have been well represented in the academic literature. To generate exhaustive reachable sets for all agents in a conflict, they have used Markov chains with kinematics-constrained transition functions [2], Bayesian probabilistic occupancy grids with linear projections of kinematic parameters [27], and data-driven learned mechanisms like higher-order control barrier functions [29] or decision trees and random forests with natural language explanations of risk [32]. While these are expected to improve accuracy compared to open-loop assumptions on agent behavior such as those quantified by the Responsibility-Sensitive Safety (RSS) metrics for AV safety [38], there are limited validation studies on real-world data with physically explainable 'why' for the human-driven 'what if' reachable sets. How and when a human driver reacts to conflict is a function of their perceived risk [40], as quantified by computational models calibrated against laboratory data [22], deep learning networks with abstracted Operational Design Domain (ODD) features [35], and collision avoidance difficulty using parameters like visual looming and minimal control effort needed [14]. Physics-based, model-in-loop counterfactual simulators to predict crash configurations [28] and simulation-based models for humans avoiding crashes are used to design more effective ADAS for nominal drivers [44] and distracted drivers [4]. Human ability to navigate complex traffic situations has been modeled using future, counterfactual tracks with collision probabilities [3]. Studies have compared real-world behavior to model behavior on a few conflict types and benchmark scene-sets [33][31]. While such sources can be used to formulate human responder models, we are not aware of any models that have been verified in terms of aggre-

gate collision risk to establish AUR and PRB. We ensure *model explainability* by formulating human-data-driven behavior models, and address the *collision benchmarking* gap by embedding them into an ADS risk framework that outputs a collision metric that can be compared to observed real-world collisions.

Accurate aggregate collision risk estimation is desired within collectible data sizes. Software releases in simulation are evaluated using leading indicators of collisions: safety surrogate metrics (SSMs) and metrics indicative of defensive yet predictable driving[43] [21], such as compliance to the state vehicle code, comfort for the rider, and minimum disruption to other road users. Such indicators have the advantage of frequent exposure compared to collisions and therefore need smaller simulation sample sizes to evaluate ADS performance [39] [1]. However, predictive correlations between SSMs and collisions need copious amounts of data and causality is difficult to prove. Thus, using SSMs alone for assessing absolute risk and establishing AUR/PRB is inadequate. To address this *metric granularity* gap, we propose a collision risk metric that is more frequent and expected to be more statistically accurate than discrete collisions, while being more defensible in traditional safety standards than SSMs or other behavioral metrics.

3. Framework Description

The proposed collision risk framework takes as inputs trajectories of any two-agent conflict in simulation, and assesses the collision risk of that conflict using a series of steps described below (summarized in Figure 2), where quality of the components are justified by referenced models that are improved upon and verified. All subsections are applicable to an open-loop simulation environment. If we use a closed-loop simulator instead, Section 3.1-3.2 can be eliminated and the remaining framework is still valid, as long as the *sim2real* gap for the simulated environment and agents is sufficiently low. The framework is verified in a synthetic simulation environment using naturalistic trajectories for the agents (Section 4), and its impact on ADS development demonstrated in a synthetic and sensor replay environment using a simulated ADS as one of the agents (Section 5).

3.1. Classification of Conflict, Initiator, Responder

Identified conflicts are first classified into one of 32 different types defined for our ODD. Our conflict taxonomy was initialized using NHTSA definitions (e.g crash type variable in CRSS/FARS coding) and academic literature (e.g., [28][34], [25]) and extended based on internal analysis on our internal ADS test data. For example, we granularize NHTSA definition of "same trafficway, same direction, angle sideswipe" further into distinct conflict types for pullouts, lane changes, short merges, cut-ins, etc. While inter-
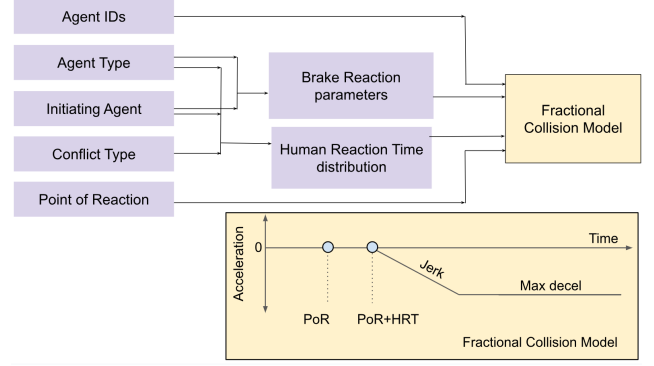

Figure 2. Summary of fractional collision workflow. Inset shows a notional reaction model for the responder in a 2-body conflict

nal efforts are underway to automate conflict classification using machine learning (ML), the model presented here relies on heuristics for automated conflict classification with appropriate quality assurance (QA) by human triage.

We define an *initiator* as the agent in a conflict that initiates a surprising behavior that the other agent (*responder*) would need to act upon to avoid propagating the conflict [25] [36]. This is not assignment of fault or blame (legal implications), and instead an automation-friendly step that ensures appropriate application of our *responder* behavior models to open-loop simulation. For example, an agent phantom braking in front of a following agent on a road with no traffic lights and no threat of crossing agents ahead, makes the first agent an initiator since they started the surprising behavior.

3.2. Behavior Models for the Responder

The Point of Reaction (PoR) is the earliest timestamp that a responder in a simulation could have identified an impending conflict. For a human responder, the PoR is the first time the human notices conflicting or surprising behavior by the initiator. PoR also represents the timestamp at which we 'start the clock' to measure human reaction time (HRT) (Figure 2-inset). Since literature-derived and proprietary HRT (Section 3.2) is very sensitive to 'start time' definitions, our models are only valid when PoR is properly defined. The definition of PoR for each conflict could be defined either from the perspective of the driver or the perspective of an omniscient third party. For the former, HRT begins when the driver has sufficient perception of the conflict, e.g., a surprise-based framework [12]. For the latter [11], the driver in that same conflict uses different information toward their decision, and may actually react sooner leading to the HRT distribution including negative values. While internal efforts are underway to automate PoR using ML, such as using the concept of modeling surprise, the models used in the paper rely on heuristics for automated

PoR with human triage QA.

We have built post-PoR responder models based on literature as well as trajectory data from other road users as extracted from our internal, proprietary logs of ADS driving. Our models follow the principles of Non-Impaired and Eyes on the conflict (NIEON) [24] when assuming intent and ability of the responder. Inattentive agents are modeled separately as a function of known populations that cause conflicts [30] to account for residual risk. Kinematic reaction parameters to model the behavior of the responder after their PoR include HRT, longitudinal and lateral jerk, and steady state acceleration [8], notionally represented in Figure 2-inset. We chose HRT as the primary parameter to model spread in our behavior models due to maximum sensitivity toward severity outcomes, and longitudinal jerk and acceleration as secondary parameters. Lateral parameters were considered in a tertiary corrective step in the QA phase of the workflow due to limited data available for an automated model within acceptable uncertainties.

Reactor models are known to be sensitive to the agent type and the conflict type [30] hence multi-variate distributions are modeled. For example, the braking capacity of a sedan, truck, and bicycle are very different. Measurement of HRT and other kinematic parameters in the incidence of imminent danger is also sensitive to the experimental setup if naturalistic data is not readily available or has noisy dependence on naturalistic parameters. For example, the HRT for rear-end collisions[30] and cut-in conflicts[11] were obtained for SHRP2-based studies. Cyclist evasive parameters were found to be different when retrieved in simulation [45] vs. laboratory environments [7] vs. field tests [45], hence a joint distribution used. Rear-end vehicular conflicts [30] were parametrized differently from vehicular cut-ins [11] and motorcyclists [16].

3.3. Loss Severity Mapping

The severity of a counterfactual crash is quantified by a crash severity model (CSM). When the two colliding objects are vehicles, the model employs a momentum-based contact algorithm to estimate the velocity differentials ($delta_v$) experienced by each vehicle (Ref [6]). If a vehicle's mass is unknown, the mass and inertia of the vehicle are estimated based on the rectangular footprint area. The initial speeds, contact location, and relative orientation are automatically input to the model from the trajectory information. The mapping of severity vs $delta_v$ approximates the L1 and L2 severity levels ascribed by Virginia Tech Transportation Institute (VTTI) in their evaluation of SHRP2 collisions. The threshold $delta_v$ between L1 and L2 was approximated to be 6 mph. Since VTTI's highest severity (L1) does not distinguish between injury scales, we introduced an additional level of severity (L0), thresholded at 20mph, to categorize higher-severity crashes.
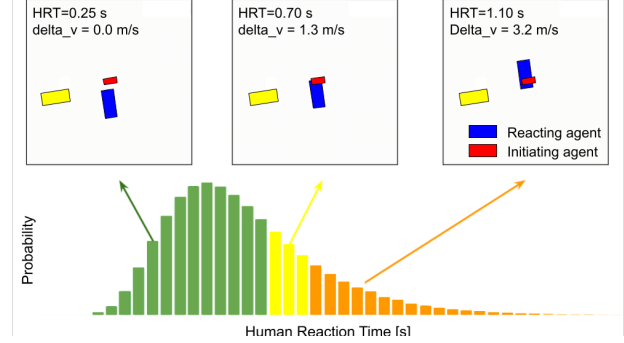

Figure 3. Example of an HRT distribution for a human responder, while holding other kinematic parameters constant in the joint probability model, used to obtain the $delta_v$ for each bin. The weighted sum across all bins is used to obtain the fractional collision score, i.e. probability of L1 (orange), L2 (yellow), no collision (green). Insets show the counterfactual behaviors that led to the arrowed bin's outcome. While a single parameter (HRT) is used for visual convenience, fractional collisions can include any number of behavioral parameters (or even extended to uncertainties associated with simulation features) whose probability mass functions are known

Crash severity between a vehicle and vulnerable road users (VRUs) are evaluated using the relative resultant contact speed at time of impact. A speed below 5mph is considered an L2 event, speeds between 5 and 15 mph are considered an L1 event, and speeds above 15 mph are considered to be L0 events. Our severity gradation is roughly analogous to standardized definitions[18]. IS0 26262 specifies an S1 (or higher) event to be at least 10% probability of AIS1 (Abbreviated Injury Scale), and an S2 (or higher) event to be at least 10% probability of AIS3. Consequently, there is a rough mapping between IS0 26262 and our severity classifications, with S0 (ISO) equivalent to L2 (proposed), S1 (ISO) equivalent to L1 (proposed) and S2 (ISO) equivalent to L0 (proposed). Alternative formulations for severity quantification verified using crash experiments and database results from Volvo [41] may also be used.

3.4. Risk Scoring as Fractional Collisions

We use the severity from each counterfactual reaction in any identified conflict to estimate the collision risk of that counterfactual. We then probabilistically aggregate the severity across all counterfactual scenarios to estimate the total collision risk for that conflict as a *fractional collision*. Fractional collisions can be numerically added across all conflicts in a given mileage sample size to assess the total collision risk in the form of total number of collisions. The outcome is thus easy to interpret and benchmark while the underlying models capture the uncertainties associated with various assumptions in the simulation.

Figure 3 illustrates the fractional collision outcome for

an example scenario across a probabilistic distribution of kinematic parameters and visualized on the HRT axis. We model the collision between a cyclist (initiator) and a vehicle (responder). After selecting an appropriate PoR (first timestamp where the cyclist is visible to the vehicle's driver and has crossed the stop line at the intersection), we simulate the brake reaction of the vehicle over its parameters. The resulting $delta_v$ for each combinatorial set of parameters is calculated and then converted to the appropriate injury severity: No collision (NC) $Lnone$ (green), property-damaging $L2$ (yellow), injury-causing $L1$ (orange), injury-causing $L0$ (red). The sum of the injury severity, weighted by the joint probability of parameters, provides the probabilistic collision value: P(NC), P(L2), P(L1), P(L2). The mass function in this example shows 20% probability of L1, 30% chance of L2, and 50% chance of NC.

Collision risk is computed as a function of the conflict initiator. Fractional collisions for ego-initiated conflicts account for the responder's probabilistic evasiveness, e.g., ADS caused 1.67 L_x total collisions in spite of expected response. Note that we also compute an ADS's conflict avoidance and compliance risk using SSMs without any assumptions on responder evasiveness. For agent-initiated conflicts, we compute two loss distributions; One for ADS response, which is a single point describing the loss severity ($L_x \in [L0, L1, L2, Lnone]$c) of the collision in simulation, and another for a modeled human quantified by our probabilistic distribution of severity (like figure 3). The difference in these distributions represents the relative risk of ADS response when compared to modeled humans for each loss severity. The aggregated risk of the ADS response across all severity can be compared to humans, e.g., ADS avoided 4.06 L_x collisions compared to expected response.

Fractional collisions can be used to evaluate the quality of ADS software. Conflicts with unreasonable collision risk are used to improve the ADS, provide feedback toward reducing the *sim2real* gap, and to improve our models for risk assessment described above. Aggregate results may also be used to assess human-like driving by comparing total collision risk to naturalistic driving datasets available from NHTSA or other government sources, with known caveats for under-reporting (e.g., 55% of all vehicles with property damage are unreported, 32% of injuries occur in unreported crashes [5]). While this paper focuses on losses associated with other road users and their pets, our collision risk framework can also be extended to nonhuman conflicts such as those with road debris, stray animals, and map features by mapping to different and appropriate loss severity.

4. Framework and Fractional Collision Eval

We assess our proposed end-to-end framework, whose components are quality assured per Section 3, against two naturalistic data sets. One data set contained agent tra-

jectories obtained from telemetry/video data of VTTI's SHRP2 dataset, collected by voluntary subject vehicles (SV), retrofitted with recording instruments and reconstructed by GCAPS (Global Center for Automotive Performance Simulation) [13]. The other data set of naturalistic agent trajectories was annotated reconstructions from dashboard camera videos provided by a commercial provider called Nexar. [26]. Alternative track-annotated sources considered were APTIV's *nuScenes* dataset [9] or commercially available drone or roadway camera footage of accident-prone locations such as high-speed merges or intersections. Since our workflow's verification is ideally performed on large mileage of naturalistic data across all ODDs driven randomly, static test data does not provide the coverage necessary for all conflicts. Our proprietary ADS data from public roads was considered inappropriate too because it is typically collected with an expert safety driver behind the wheel (with faster-than-average ability to perceive and react) and, when collected with the ADS in closed loop, is interrupted by ADS disengagements (which causes the part of the scene we are most interested to rarely play out naturally).

Queried from nearly 8 million miles of SHRP2 data, 33 collisions and 497 near-collisions (NC) were found among all conflicts on surface streets involving two agents and excluding rear-end struck, animal, or inanimate object conflicts (since an evasive maneuver is not expected). Near-collisions were further restricted by those where the non-SV was at-fault since SHRP2's NC are identified based on the SV's evasive maneuver. To verify our proposed framework, we selected 17 collisions (split equally where the SV is initiator/responder) and 65 non-collisions to be qualitatively representative of the expected population distribution of possible responders for diverse scenario coverage. From an unverifiable mileage of Nexar video, 190 collisions and 44 NC were procured as potential reconstructions. The workflow described in Section 3 was applied to the selected 316 events and fractional collisions computed as probabilities of L_x. They are compared against the ground truth (GT) outcome of each scene, i.e. discrete L_x using annotated outcome cross-checked against our CSM (Section 3.3) applied to the conflict trajectories, as well as to a no-reaction-model (NRM), i.e. a naive control model where no deceleration is allowed after PoR.

Each scene includes a GT responder should represent a datapoint on our model's joint distributions (i.e. human population), thus some scenes may have higher severity in GT than expected while others may have lower severity in GT than expected. We also expect fewer fractional collisions than GT and NRM for scenes with collisions and more fractional collisions and NRM collisions than GT for scenes with NC. In fact, NRM may be a datapoint outside the NIEON distribution. We *hypothesize* that when *aggregated*

across an unbiased sample of mileage across all ODDs and conflicts that a vehicle encounters, total fractional collisions should closely match GT more than NRM does. This aggregate hypothesis should also hold for a set of scenarios that represent proportionately the human population of responders. The proposed framework is a powerful predictor of collision risk in a dataset while accounting for population variability in *open-loop simulations* where the logged agents cannot/do not respond to the simulated ADS.

4.1. Scene-by-Scene Verification

For every scene, we implement three basic checks to ensure data quality of the reconstructions and verify that our workflow design and execution matches our intent. First, if the GT scene has an annotated collision or NC, our reconstruction of that scene through severity mapping (Section 3.3) should output as such. Second, the fractional collision outcome (Section 3.4) should not show zero probability for a severity that corresponds to the original severity. Third, the severity outcome of the NRM should be equal to or worse than GT because humans are expected to react appropriately to *reduce* severity of an impending collision. 81 of 82 SHRP2 scenes were successfully reconstructed in our synthetic sim environment and passed all checks. The one exception where human reaction worsens the collision was used to improve evasive maneuver models. The Nexar-based reconstructions, owing to greater scenario volume and coverage of more conflict types and agents, surfaced need for both data quality and model improvements. Of 239 scenes, 17% were found to be poorly reconstructed in the simulator due to sudden jerk associated with the original collision being extrapolated unrealistically for longer HRTs or NRM. Of the others, ∼10% failed one of the three basic checks due to framework-intrinsic reasons such as vehicular polygons in the reconstructions being underestimated, PoR being inaccurate, and original human driver being fully nonreactive (violating NIEON).

Our probabilistic framework is more accurate in predicting the most likely outcome but also ensures full recall of worst outcomes in any scene. The most probable severity predicted by fractional collisions, i.e. the severity L_x corresponding to $max[P(L_x)]$, matched the GT severity for 91% of the scenes. The maximum fractional collision, i.e. the most severe $L_x \forall P(L_x) \geqslant 0$, matched the GT severity in only 62% of the scenes, however was lower than GT severity in *all* scenes. The most probable severity and maximum fractional collision matched GT in 51% and 71% Nexar scenes, respectively, and surfaced GT severity in *all* scenes.

4.2. Aggregate Verification

SHRP2 and Nexar complement each other to provide quantitative aggregates and ODD/agent coverage with explainable trends, respectively, to verify framework effectiveness.

While the SHRP2 scenes were selected such that the responder spread was representative of the human population, we could not inform selection proportions with Nexar since video content was proprietary before procurement. To verify our aggregate hypothesis and as summarized in Table 1, we compare the sum total of fractional collisions computed from the synthetic simulations of all SHRP2 and Nexar scenes to the actual GT collisions in those scenes and to the number of discrete collisions that a NRM would had in those scenes, organized by whether there was a GT collision (rows) and by severity of the collisions (columns).

Compared to the 17 collision scenes from SHRP2, our framework estimated 12.6 collisions, overestimating $L0$ and $L1$ but underestimating $L2$ (shaded green). In the 65 NC, i.e. severity $Lnone$, our framework estimated 4.56 fractional collisions (pink). However, when aggregated across all scenes and severity, the total fractional collisions is 17.15, only 1% over the actual 17 and 2x closer than using the NRM which predicts 34 collisions (blue). As expected, the NRM predicts a collision in every collision-scene, but the 17 collisions in NRM are far more skewed at higher severity (yellow) than the GT's lower severity (green). NRM shows 9 L0 collisions while there were none in GT (also yellow).

Nexar-based aggregate results show 40% fractional collisions compared to all GT collisions. This can be partially attributed to the disproportionately high ratio of collisions (131) to NC (37) in the procured data. In comparison, our SHRP2 selections had a C:NC ratio 17:65 and the SHRP2 database for ODD-conflicts was at 1:15. While the high C:NC ratio procured was motivated to maximize difficult scenes, Nexar's 50x C:NC ratio compared to unbiased mileage coupled with the fact that fractional collisions will score lower than discrete collisions (=1) for any GT scene, sets the aggregate up for under-prediction. Therefore, Table 1 shows lower fractional collisions for every severity level (shaded green). NRM over-predicts GT as expected (shaded yellow), but by only 8% compared to 100% in SHRP2 (blue). To predict *total* collisions accurately, the dataset must contain sufficient near-collision scenarios.

5. Framework Utility in ADS Behavior Eval

Our verified framework can be used to evaluate an arbitrary ADS software release (Fig 1-C as ADS) in simulation environments (Fig 1-A) to demonstrate the impact of collision risk outcomes predicted. The simulation environment can be of two types: (i) a high-volume re-simulation of logged sensor data, and (ii) a high-volume synthetic simulation of imagined, virtual environments and agents - both for Fig 1-B. The ADS trajectory is the output of our proprietary simulator that plays out the ADS onboard software in a virtual environment as it 'drives' through perception inputs that are either synthetic or open-loop replayed from sensor

6

Post-PoR Framework	GT Collision?	SHRP2					Nexar				
		Severity					Severity				
		L0	L1	L2	NC	Total	L0	L1	L2	NC	Total
Fractional Collisions	Yes	1.16	10.31	1.13	4.40	12.60	8.75	32.82	30.77	58.53	72.33
	No	0.92	1.61	2.03	60.41	4.56	0.00	1.11	5.29	30.54	6.41
	All	2.08	11.92	3.16	64.81	17.15	8.75	33.93	36.07	89.07	78.74
Ground Truth (GT)	Yes	0	10	7	0	17	11	61	59	0	131
	No	0	0	0	65	0	0	0	0	37	0
	All	0	10	7	65	17	11	61	59	37	131
No Reaction Model (NRM)	Yes	2	13	2	0	17	20	73	38	0	131
	No	7	6	4	48	17	1	3	6	27	10
	All	9	19	6	48	34	21	76	44	27	141

Table 1. Collision severity for selected SHRP2 and Nexar scenarios as a function of the framework used to model the post-PoR behavior of the responder, further split by whether there was an actual collision on-road (GT). Colors are explained in the main text narrative.
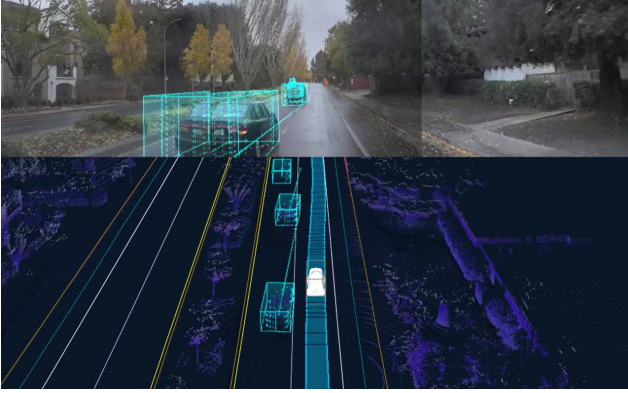

Figure 4. Example of an agent-initiated conflict with a non-zero risk of simulated collision. The other agent accelerates while changing lanes into the ADS, which nudges away and slows down but does not avoid collision. Screenshot marks the point of reaction of the ADS. The fractional collision results indicate that 56% of human responders would be able to avoid this collision.

logs. The simulator latency models are trained (and validated) using onroad and hardware-in-loop data so that each module's latency distributions are representative of the real world. Our framework (Fig1-D) enables surfacing ADS issues, estimates collision risk in ADS-initiated conflicts and collision risk eliminated by ADS in agent-initiated conflicts, and supports policy discussions on AUR and PRB.

5.1. Agent-Initiated Conflicts

Replay simulation is an important mode of sourcing realistic agent-initiated conflicts because track overlaps are more frequent than in onroad testing. This is due to the pose divergence artifact described in Section 2, causing the initiator (non-ADS) to 'behave' unrealistically aggressively. Fractional collisions in such scenarios can surface agent-initiated risk and realistically quantifying PRB per scene. 376 successfully simulated agent-initiated conflicts were identified in a quarter-million mile *log simulation*. Of these, only 13 showed a fractional collision severity that was better than the severity of the ADS's simulated outcome. The PRB gap in 3.5% scenes is a quantitative measure of the ADS performance, and the 13 examples are opportunities for autonomy improvement because some humans could outperform the ADS (by either mitigating/lowering the severity of collision or avoiding the collision entirely). Figure 4 shows an example of a conflict where the ADS is driving on a two-lane road at 35 mph. The vehicle in the left lane cuts into the lane of the ADS; due to pose divergence, this cut-in occurs approximately adjacent to the ADS. The simulated ADS vehicle nudges away from the initiator and decelerates moderately but still has an L2 collision. The modeled human shows fractional collisions of 6% L1, 38% L2, and 56% no collision. Since more than half (56%) of the humans modeled performed better than the ADS, this scene is an opportunity for autonomy improvement.

205 scenes with assured data quality were selected from the SHRP2 and Nexar sets in Section 4, reconstructed in our proprietary *synthetic simulation* environment, and simulated with the responder replaced with our simulated ADS. The simulated ADS resulted in 26 collisions, compared to 103 collisions in the original human outcome (i.e. ADS reduced actual collisions by 4x), and 68.42 fractional collisions by the modeled human (i.e. ADS reduced expected collisions by 2.6x). In the scenes that ADS simulated to a collision, only 4 had fractional collisions better than the ADS loss. Thus, 97% of a set of handpicked hardest scenarios sourced from tens of millions of miles of naturalistic driving confirmed that a human could not improve collision risk compared to ADS. Our framework establishes PRB for the vast majority of agent-initiated conflicts and enables the
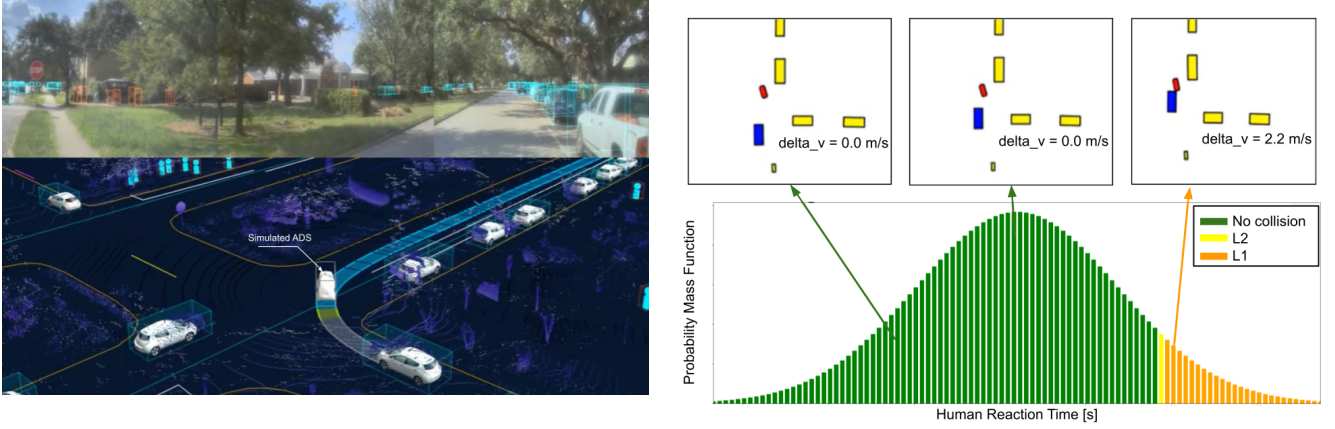
Figure 5. [Left] Example of an ADS-initiated conflict with a non-zero risk of simulated collision. The ADS unnecessarily yields during an unprotected right turn. Screenshot marks the PoR by the responder. The fractional collision results indicate that 91.5% of human responders would be able to avoid this collision. [Right] Probability mass function associated with the HRT distribution for this conflict type for a particular combination of other probabilistic kinematic parameters constant, which is used to compute fractional collisions

automated identification of rare agent-initiated issues to fix.

5.2. ADS-Initiated Conflicts

The detection of ADS-initiated conflicts in any simulation (*replay or synthetic*) and using fractional collisions to quantify risk is an important means of must-fix issue discovery, especially to establish AUR. The vast majority of our synthetic simulations are aimed at providing coverage extended beyond replay sim mileage for ADS-initiated conflicts, and our proposed framework can identify issues from their outcomes based on collision risk (in parallel to SSMs or other metrics). Fractional collisions increases our issue discovery signal. From the quarter-million mile log simulation, six conflicts involving the ADS and another vehicle were found with non-zero fractional collisions. These can be numaerically added to estimate the expected quantity of losses of each severity in the said mileage: the total fractional collisions were 0 L0, 0.4 L1, 1.7 L2. Thus, our framework surfaced 6 scenarios with injury-causing risk despite no collision being likely in that mileage. As a concrete example of an individual scenario, Figure 5 shows an example conflict resulting from an unnecessary yield by the ADS while making an unprotected right turn into traffic. Cross-traffic had been yielding to a crossing school bus, providing the ADS an opportunity to make the turn; however, the ADS then unnecessarily stops after entering the lane. The responder does not react to the ADS in simulation due to pose divergence, and thus the scenario was flagged as a conflict. Our framework showed that the conflict had probabilities of 8.5% for L1 and 91.5% for no collision. Note that pose divergence in log replays can cause unrealistic responder behavior between PoR and HRT onset (for example, acceleration), thereby inflating risk. Since ADS should not initiate collision-risking conflicts and our goal is to discover must-

fix issues, this inflation is accepted to prioritize recall.

6. Future Work

Improvements expected in future work can be categorized as [1] intrinsic improvements to our proposed framework, [2] extending the framework to add uncertainties associated with the simulation environment and agents, and [3] expanding the naturalistic data set to test our updated framework on. For the first, we have pointed out work underway for various components in Section 3, especially those identified in Section 4 (e.g. PoR). Further development and verification of agent models is required. For example, to accurately model cyclist responses, evasive maneuvers other than simple brake responses (e.g., nudging) must be considered. Simulations in Section 5.2 found six conflicts that were possible ADS-initiated collisions involving a cyclist, where the current model appeared to overestimate the likelihood of an injury-causing collision by assuming that no responder would deviate from the logged trajectory when a slight nudge may have been sufficient to avoid the collision. Since the SHRP2/Nexar datasets did not contain cyclist responders, more data will be needed to verify the new models. For the second, we will identify the most sensitive parameters in the simulation environment that affect *sim2real* gaps and collision risk when compared to onroad data, and model them within the proposed framework. Fractional collisions can thus account for some of the uncertainties associated with the tools used to evaluate them. This is especially relevant as simulators improve their capabilities in sensor modeling, neural reconstructions, world-model controlled agents, interaction sensitivity to latency, ADS controls models, etc. For the third, we will improve the quality of reconstructions, and explore sources of naturalistic data that improves representativeness of the human population and verification of the framework.

References

[1] Abdul Razak Alozi and Mohamed Hussein. Active road user interactions with autonomous vehicles: Proactive safety assessment. *Transportation research record*, 2677(7):74–89, 2023. 3

[2] Matthias Althoff, Olaf Stursberg, and Martin Buss. Safety assessment of autonomous cars using verification techniques. In *2007 American Control Conference*, pages 4154–4159, 2007. 2

[3] Stefan Annell, Alexander Gratne, and Lars Svensson. Probabilistic collision estimation system for autonomous vehicles. In *Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016. 2

[4] Jonas Bärgman, Malin Svärd, Simon Lundell, and Erik Hartelius. Methodological challenges of scenario generation validation: a rear-end crash-causation model for virtual safety assessment. *Transportation Research Part F: Traffic Psychology and Behaviour*, 104:374–410, 2024. 2

[5] Lawrence Blincoe, Ted R Miller, Jing-Shiarn Wang, David Swedler, Tom Coughlin, Bruce Lawrence, Feng Guo, Sheila Klauer, and Thomas Dingus. The economic and societal impact of motor vehicle crashes, 2019. Technical Report DOT-HS-813403, National Highway Traffic Safety Administration, 2022. 5

[6] Raymond M Brach, Kevin J Welsh, and R Matthew Brach. Residual crush energy partitioning, normal and tangential energy losses. *Crash Reconstruction Research*, 138:45, 2008. 4

[7] Kim Buchholtz and TL Burgess. An evaluation of bicycle-specific agility and reaction times in mountain bikers and road cyclists. *South African journal of sports medicine*, 32 (1):1–5, 2020. 4

[8] Jonas Bärgman, Christian-Nils Boda, and Marco Dozza. Counterfactual simulations applied to shrp2 crashes: The effect of driver behavior models on safety benefit estimations of intelligent safety systems. *Accident Analysis and Prevention*, 102:165–180, 2017. 4

[9] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5

[10] M.L. Cummings and Ben Bauchwitz. Identifying research gaps through self-driving car data analysis. *IEEE Transactions on Intelligent Vehicles*, pages 1–10, 2024. 1, 2

[11] Swaroop Dinakar, Jeffrey W. Muttart, Darlene E. Edewaard, Michael Giannone, and Connor Dickson. Driver response time in cut-off scenarios from the second strategic highway research program naturalistic database. *Transportation Research Record*, 2676(2):706–717, 2022. 3, 4

[12] Johan Engström, Shu-Yuan Liu, Azadeh Dinparastdjadid, and Camelia Simoiu. Modeling road user response timing in naturalistic traffic conflicts: A surprise-based framework. *Accident Analysis and Prevention*, 198:107460, 2024. 3

[13] Jonathan M Hankey, Miguel A Perez, and Julie A McClafferty. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Technical report, Virginia Tech Transportation Institute, 2016. 5

[14] Xiaolin He, Riender Happee, and Meng Wang. A new computational perceived risk model for automated vehicles based on potential collision avoidance difficulty (pcad). *Transportation Research Part C*, 166(104751), 2024. 2

[15] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[16] Pedro Huertas-Leyva, Marilee Nugent, Giovanni Savino, Marco Pierini, Niccolò Baldanzini, and Simon Rosalie. Emergency braking performance of motorcycle riders: skill identification in a real-life perception-action task designed for training purposes. *Transportation research part F: traffic psychology and behaviour*, 63:93–107, 2019. 4

[17] ISO/PAS 21448:2022(en). Road vehicles — safety of the intended functionality. Standard, International Organization for Standardization, Geneva, CH, 2022. 1

[18] ISO/PAS 26262:2018(en). Road vehicles — functional safety. Standard, International Organization for Standardization, Geneva, CH, 2018. 4

[19] ISO/PAS 8800:2024(en). Road vehicles — safety and artificial intelligence. Standard, International Organization for Standardization, Geneva, CH, 2024. 1

[20] Nidhi Kalra and Susan M. Paddock. 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, 2016. 1

[21] Hoseon Kim, Jieun Ko, Cheol Oh, and Seoungbum Kim. Evaluation of autonomous driving safety by operational design domains (odd) in mixed traffic. *Sustainability*, 16(22): 9672, 2024. 3

[22] Sarvesh Kolekar, Joost De Winter, and David Abbink. Human-like driving behaviour emerges from a risk-based driver model. *Nature communications*, 11(1):1–13, 2020. 2

[23] Philip Koopman, William H Widen, et al. Breaking the tyranny of net risk metrics for automated vehicle safety. *Safety-Critical Systems eJournal*, 3(1), 2024. 1

[24] Kristofer D. Kusano, Kurt Beatty, Scott Schnelle, Francesca Favaro, Cam Crary, and Trent Victor. Collision avoidance testing of the waymo automated driving system, 2022. 4

[25] Kristofer D Kusano, John M Scanlon, Mattias Brännström, Johan Engström, and Trent Victor. Framework for a conflict typology including contributing factors for use in ads safety evaluation. In *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, number 23-0328, 2023. 3

[26] Lev Yitzhak Lavy. Modular vehicle sensing, assisting connected system, 2022. US Patent 11,405,235. 5

[27] Philippe Ledent, Anshul Paigwar, Alessandro Renzaglia, Radu Mateescu, and Christian Laugier. Formal validation of probabilistic collision risk estimation for autonomous driving. In *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, pages 433–438, 2019. 2

[28] Alexandros Leledakis, Magdalena Lindman, Jonas Östh, Linus Wågström, Johan Davidsson, and Lotta Jakobsson. A method for predicting crash configurations using counterfactual simulations and real-world data. *Accident Analysis and Prevention*, 150:105932, 2021. 2, 3

[29] Karen Leung, Sushant Veer, Edward Schmerling, and Marco Pavone. Learning autonomous vehicle safety concepts from demonstrations. In *2023 American Control Conference (ACC)*, pages 3193–3200. IEEE, 2023. 2

[30] Gustav Markkula, Johan Engström, Johan Lodin, Jonas Bärgman, and Trent Victor. A farewell to brake reaction times? kinematics-dependent brake response in naturalistic rear-end emergencies. *Accident Analysis and Prevention*, 95: 209–226, 2016. 4

[31] Konstantinos Mattas, Michail Makridis, George Botzoris, Akos Kriston, Fabrizio Minarini, Basil Papadopoulos, Fabrizio Re, Greger Rognelund, and Biagio Ciuffo. Fuzzy surrogate safety metrics for real-time assessment of rear-end collision risk. a study based on empirical observations. *Accident Analysis and Prevention*, 148:105794, 2020. 2

[32] Richa Nahata, Daniel Omeiza, Rhys Howard, and Lars Kunze. Assessing and explaining collision risk in dynamic environments for autonomous driving safety. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pages 223–230. IEEE, 2021. 2

[33] Pierluigi Olleja, Gustav Markkula, and Jonas Bärgman. Validation of human benchmark models for automated driving system approval: How competent and careful are they really? *Accident Analysis and Prevention*, 213:107922, 2025. 2

[34] Martin Östling, Nils Lubbe, Hanna Jeppsson, and Pradeep Puthan. Passenger car safety beyond adas: defining remaining accident configurations as future priorities. In *The 26th International Technical Conference on the Enhanced Safety of Vehicles, Eindhoven, Netherlands*, pages 19–91, 2019. 3

[35] Peng Ping, Yuan Sheng, Wenhu Qin, Chiyomi Miyajima, and Kazuya Takeda. Modeling driver risk perception on city roads using deep learning. *IEEE Access*, 6:68850–68866, 2018. 2

[36] John M. Scanlon, Kristofer D. Kusano, Tom Daniel, Christopher Alderson, Alexander Ogle, and Trent Victor. Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. *Accident Analysis and Prevention*, 163:106454, 2021. 2, 3

[37] Matthew Schwall, Tom Daniel, Trent Victor, Francesca Favaro, and Henning Hohnhold. Waymo public road safety performance data. *arXiv preprint arXiv:2011.00038*, 2020. 1

[38] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*, 2017. 2

[39] Amolika Sinha, Sai Chand, Kasun P. Wijayaratna, Navreet Virdi, and Vinayak Dixit. Comprehensive safety assessment in mixed fleets with connected and automated vehicles: A crash severity and rate evaluation of conventional vehicles. *Accident Analysis and Prevention*, 142:105567, 2020. 3

[40] Heikki Summala. Risk control is not risk adjustment: the zero-risk theory of driver behaviour and its implications. *Ergonomics*, 31(4):491–506, 1988. 2

[41] Linus Wågström, Alexandros Leledakis, Jonas Östh, Magdalena Lindman, and Lotta Jakobsson. Integrated safety: Establishing links for a comprehensive virtual tool chain. In *26th International Technical Conference on the Enhanced Safety of Vehicles (ESV): Technology: Enabling a Safer Tomorrow, Eindhoven, the Netherlands*, 2019. 4

[42] Jun Wang, Li Zhang, Yanjun Huang, and Jian Zhao. Safety of autonomous vehicles. *Journal of advanced transportation*, 2020(1):8867757, 2020. 1

[43] Lukas Westhofen, Christian Neurohr, Tjark Koopmann, Martin Butz, Barbara Schütt, Fabian Utesch, Birte Neurohr, Christian Gutenkunst, and Eckard Böde. Criticality metrics for automated driving: A review and suitability analysis of the state of the art. *Archives of Computational Methods in Engineering*, 30(1):1–35, 2023. 3

[44] Peter Wimmer, O Op_Den_Camp, Hendrik Weber, Henri Chajmowicz, Michael Wagner, Jorge Lorente Mallada, Felix Fahrenkrog, and Florian Denk. Harmonized approaches for baseline creation in prospective safety performance assessment of driving automation systems. In *27th international technical conference on the enhanced safety of vehicles (esv), yokohama, japan*, pages 3–6, 2023. 2

[45] Jing Xu, Yingzi Lin, and David Schmidt. Exploring the influence of simulated road environments on cyclist behavior. *International Journal of Virtual Reality*, 17(3):15–26, 2017. 4

[46] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 2

[47] Rui Zhou, Guoqing Zhang, Helai Huang, Zhiyuan Wei, Hanchu Zhou, Jieling Jin, Fangrong Chang, and Jiguang Chen. How would autonomous vehicles behave in real-world crash scenarios? *Accident Analysis and Prevention*, 202: 107572, 2024. 2