Building Safer Autonomous Agents by Leveraging Risky Driving Behavior Knowledge

Ashish Rana  
Thapar Institute of Eng. & Tech., India  
aranabe15@thapar.edu

Avleen Malhi  
Bournemouth University, UK  
amalhi@bournemouth.ac.uk

Abstract—Simulation environments are good for learning different driving tasks like lane changing, parking or handling intersections etc. in an abstract manner. However, these simulation environments often restrict themselves to operate under conservative interaction behavior amongst different vehicles. But, as we know, real driving tasks often involve very high risk scenarios where other drivers often don’t behave in the expected sense. There can be many reasons for this behavior like being tired or inexperienced. The simulation environment doesn’t take this information into account while training the navigation agent. Therefore, in this study we especially focus on systematically creating these risk prone scenarios with heavy traffic and unexpected random behavior for creating better model-free learning agents. We generate multiple autonomous driving scenarios by creating new custom Markov Decision Process (MDP) environment iterations in the highway-env simulation package. The behavior policy is learnt by agents trained with the help from deep reinforcement learning models. Our behavior policy is deliberated to handle collisions and risky randomized driver behavior. We train model free learning agents with supplement information of risk prone driving scenarios and compare their performance with baseline agents. Finally, we casually measure the impact of adding these perturbations in the training process to precisely account for the performance improvement obtained from utilizing the learnings from these scenarios.

Index Terms—Autonomous Agents, Driving Simulations, Trajectory Prediction, Causality

I. INTRODUCTION

The arrival of autonomous driving agents have had a great impact on the automobile industry. And it will be responsible for shaping the future of this industry as well. The current industry trend progression demonstrates that the connected fleet of autonomous agents will be dominating our driving infrastructure [1]. But, that vision is still quite far considering the safety infrastructure, security and public policy reasons [2]–[5]. Therefore, our immediate focus should be on making our driving agents safe and efficacious. Creating safer agents has been explored well for the past few years [6]–[12]. As it is crucial to know the expected agent behavior in different environments especially safety critical ones. In natural driving environments risk prone scenarios don’t happen frequently which makes the learning process harder from these scenarios [13], [14]. And it is also unethical to create these real risky driving scenarios for experimentation purposes. Therefore, generating and studying these scenarios systematically is a daunting task.

Perception systems with adversarial approaches of noisy labelling do demonstrate promising path [15]–[17]. But the underlying fundamental problem remains focused around safe interactions with other vehicles which themselves are operating independently. Simulation environments with appropriate system dynamics design assumptions do overcome this expressed safety issue. Along with that these environments allow us to systematically study the risk prone scenarios with near realistic driving behaviors. For this study we have used the highway-env simulation package which allows us to simulate different driving tasks [18]. It also provides simple interfacing capabilities to modify these environments and quickly create baseline prototypes [19] on it. We formulate the dynamics of the simulation system as Markov Decision Process (MDP) in our experimentation [20]. We model our value function approximation agents with deep reinforcement learning for these defined MDP systems [21], [22] for our study.

For our experiment design we create two distinct variants of each model architecture for above stated simulation dynamics. One of these model variants is trained with our increased dangerous driving behavior interventions for all the driving related tasks present in highway-env package. Second model
variant is the control variable in our experiment which is used to define the reward baseline for the trained agents on regular simulations. In our study we do a methodological doping of these environments with randomized dangerous driving scenarios to create more risk prone environments for driving. This is done specifically in two ways in our study, first by increasing the traffic at strategically important locations challenging the agent to make dangerous overtakes. Second, we increase the randomization factor and the clogged lanes make the environment more collision prone for our agent (ego-vehicle). This is done to create more robust agents in a given environment for that particular dynamic scenario. Which are essentially better at post impact trajectory predictions of other interacting vehicles. Figure 1 explains our causal analysis experimentation setup in sequential form.

We also attempt to understand our experimentation setup from a causal standpoint. We hold complete control of the data generation process in our experimentation. Which allows us to have a unique vantage point of conducting an experiment study which is equivalent to a randomized control trial (RCT). We train our agents with a good enough assumption of absence in unobservable confounding variables. As we have strictly defined the state dynamics and vehicle behavior governing models. Our customization capabilities in the highway-env package allows us to keep every condition the same while training except our collision scenario perturbations. Meaning that treatment and control groups are same in all aspects except our treatment i.e. there is a comparability and co-variate balance in our experimental setup [23]–[26]. With this special relation establishment we can imply that association found in our experiment setup is causation. As shown in Figure 2 our treatment (T) is subjecting the agent learning process to risk prone interacting vehicle dynamics in a given environment. After that our sample test experiment population involves evaluating the two model variants against regular and perturbed environments with excessive risk prone scenario doping. And finally by using expectation equations derived from the above causal graph we estimate the causal effect our novel learning changes for enhanced safety.

Our contributions by the means of this paper include providing benchmarking environment simulations for collision robustness predictions. Along with that we provide experimentation methodology that creates more robust agents that provide better on-road safety. And we also causally measure the impact of our risky driving behavior doping interventions for different driving environments. For the remaining paper we first discuss related work corresponding to utilizing risk prone behavior for creating safer autonomous vehicles. Second, we formally define our problem and elaborate on causal aspects of it. Third, we explain the experiment setup for creating these robust agents and elaborate upon our results. Finally, we provide the conclusion to our autonomous driving agent study.

II. PREVIOUS WORK

Deep Reinforcement Learning has been used extensively for traffic control tasks [27], [28]. The simulated environment provided by CARLA [29] gives a framework for systems that estimates several affordances from sensors in a simulated environment [30]. Navigation tasks like merging traffic requiring good exploration capabilities have also shown promising results in simulators [31]. ChauffeurNet [32] elaborates on the idea of imitation learning for training robust autonomous agents that leverages worst case scenarios in the form of realistic perturbations. A clustering based collision case generation study [33] systematically defines and generates the different types of collisions for effectively identifying valuable cases for agent training.

The highway-env package specifically focuses on designing safe operational policies for large-scale non-linear stochastic autonomous driving systems [20]. This environment has been extensively studied and used for modelling different variants of MDP, for example: finite MDP, constraint-MDP and budgeted-MDP (BMDP) [34]. BMDP is a variant of MDP which makes sure that risk notion implemented as cost signal stays below a certain adjustable threshold [35]. The problem formalization for workings of vehicle kinematics, temporal abstraction, partial observability and reward hypothesis has been studied extensively as well [34]. Robust optimization planning has been studied in the past for finite MDP systems with uncertain parameters [36]–[38] and it also has shown promising results under conservative driving behavior. For BMDP, efficacy and safety analysis has been extended into continuous kinematics states and unknown human behavior from the existing known dynamics and finite state space [35]. Model free learning networks that approximate value function for these MDPs like Deep Q-Learning (DQN) and Dueling Deep Q-Learning (DQN) networks have demonstrated promising results in continuous agent learning [39], [40].

Worst case scenario knowledge in traffic analysis has been leveraged for model based algorithms by building regions of high confidence containing true dynamics with high probability. Tree based planning algorithms were used to achieve robust stabilisation and mini-max control with generic costs. These studies also leveraged non-asymptotic linear regression
and interval prediction [20], [34] for safer trajectory predictions. Behavior guided action studies which use proximity graphs and stability trajectory computes for working with aggressive and conservative have shown promising results [41] as well. This study used C-Metric measure [22] for generating varying levels of aggressiveness in traffic. Whereas in our case we have used more randomization and traffic clogging at key areas for risk prone scenarios to measure more granular observation results.

Causal modelling techniques have contributed a lot in terms of providing great interpretative explanations in many domains [43]–[46]. Also, Fischer’s randomized control trials (RCTs) have served as the gold standard for causal discovery from observational data [23]. Sewall’s path diagrams were the first attempt of generating causal answers with mathematics [47]. Now, causal diagrams and different adjustments on these diagrams do offer direct causal relation information about any experimental variables under study [23], [24], [48], [49]. In our experimental study we use these existing mathematical tools to draw direct causal conclusions of our learnings from our environment interventions.

III. PROBLEM FORMULATION

Our goal is to design and build agents on collision prone MDPs for navigation tasks across the different traffic scenarios. The MDP comprises a behavior policy \( \pi(a | s) \) that outputs action \( a \) for a given state \( s \). With this learnt policy our goal is to predict discrete safe and efficient action from finite action set (left, right, break, accelerate, idle) for next time step for given driving scenarios.

The simulation platform that we used is compatible with OpenAI gym package [50]. The highway-env package provides the traffic flow which is governed by Intelligent Driver Model (IDM) [51] for linear acceleration & MOBIL model [52] for lane changing. MOBIL model primarily consists of safety criterion and incentive criterion. First safety criterion checks whether after lane change the given vehicle is having enough acceleration space and second criterion determines the total advantage of lane change in terms of total acceleration gain.

Given MDP \( M \) is defined as a set of \( (S, A, T, r) \) where action \( a \in A \), state \( s \in S \), reward function \( r \in [0,1]^{S \times A} \) and state transition probabilities \( T(s' | s, a) \in M(S)^{S \times A} \). We search the the behavior policy \( \pi(s|a) \) that helps us in navigating across the traffic environments to gather maximum discounted reward \( R \). The state-action value function \( Q^\pi(s,a) \) for given \( (s,a) \) assists in estimating future rewards of given behavior policy \( \pi \). Therefore, the optimal state-action value function \( Q^*(s,a) \) provides maximum value estimates for all \( s \in S \) and is evaluated by solving Bellman Equation [53], stated below for reference. From this the optimal policy \( \pi \) is expressed as \( \pi(s) = \arg \max_{a \in A} Q^*(s,a) \).

We used DQN with duelling network architecture having value & advantage streams to approximate the state-action value function \( Q(s,a) \) which predicts best possible action as learned from the policy \( \pi \).

\[
Q^*(s, a) = \mathbb{E}[R(s, a) + \gamma \sum_{s'} P(s' | s, a) \max_{a'} Q^*(s', a')] \quad (1)
\]

From our experimentation setup we intend to derive direct causal impact of our interventions in traffic environment scenarios. And as we can refer back from Figure 2, our treatment (T) is subjecting the learning agent to more risky driving behavior. Our testing sample set involves random agent reward calculation against perturbed and control environments which makes our experiment equivalent to RCT. Meaning that there is no unobserved confounding present in our experimentation i.e. backdoor criterion is satisfied. Also, in RCTs distribution of all co-variates are same except the treatment. Co-variate balance in observational data also implies that association is equal to causation while calculating the potential outcomes, refer equation stated below.

\[
P(X | T = 1) \equiv P(X | T = 0) \\
P(X | T = 1) \equiv P(X), T \perp \perp X \\
P(X | T = 0) \equiv P(X), T \perp \perp X
\]

Essentially meaning that we can use the associative difference quantity to infer the effect of treatment on outcomes. Meaning that we can use the Average Treatment Effect (ATE) approach for calculating the causal effect by simply subtracting the averaged out values treatment and control potential outcomes. In below stated equations \( Y(1) \equiv Y \mid _{T=1} \& Y(0) \equiv Y \mid _{T=0} \) and these equations hold true in case of RCTs where causal difference can be calculated with associated difference.

\[
E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)] \\
E[Y'1] - E[Y0] = E[Y \mid T = 1] - E[Y \mid T = 0]
\]

Model free learning approaches generally don’t have explicit information about the dynamics of the systems. Therefore, during the training these agents will generalize their policies corresponding to the particular given scenarios only. We introduce risky randomized behavior in these environment vehicles with collision prone randomized behavior. Figure 5 shows different probable real life collisions simulated in different highway-env package environments. It increases generalizations on less common but highly critical scenarios which can save users from hefty collisions. We critically analyze the performance of our treated agents in comparison to control agents for important tasks like performing roundabouts, handling intersections, u-turn & two-way traffic overtakes and lane changing environments in our study.

The collision between two vehicles is equivalent to intersection of two polygons in the rendered environment output of highway-env. And we detect these collisions between rectangular polygons with the separating axis theorem for the given two convex polygons. Essentially, the idea is to find a line that
Fig. 3. Possibly realistic collision examples simulated from the different driving tasks in highway-env package environments.

Fig. 4. Overlapping (Left) and Non-Overlapping (Right) projection analysis for collision detection in highway-env package.

separates both polygons. If that line exists then polygons are separated and collision hasn’t happened yet. Algorithmically for each edge of our base rectangular polygon we find a perpendicular axis to current edges under review. After that we project these edges onto that axis and in case these projections don’t overlap it functionally means no collision as rectangular polygons are not intersecting, refer Figure 4.

IV. EXPERIMENT SETUP

In our experimentation setup we calculate the ATE metric for namely lane changing, two-way traffic, roundabout, intersection and u-turn tasks, refer Figure 4. These five tasks are evaluated against increasing traffic from default vehicle count/density to a 200% increase. For each of these traffic scenarios we create our treatment environment with varying degrees of acceleration & steering parameters which would not comply with MOBIL model criteria of safety and incentives. This randomization behavior is governed by equations stated below which lays the foundation of simulating collision prone behavior. We create collision prone behavior by significantly changing the equation max-min acceleration parameters in create_random() function defined in the kinematics rules of the highway-env package. With our experimentation setup we quantify the causal model performance improvements from introduction of this risk factor knowledge in the agent learning process. And compare it with the control baseline models across these five different navigation tasks against a spectrum of increasing traffic density.

\[
\begin{align*}
acc_p &= acc_{\text{min}} + rand[0, 1] \times (acc_{\text{max}} - acc_{\text{min}}) \\
str_p &= str_{\text{min}} + rand[0, 1] \times (str_{\text{max}} - str_{\text{min}})
\end{align*}
\]  

(4)

We rebuild the highway-env package environments with our custom changes by altering the environment configurations, randomizing behavior and adding new vehicles to strategically increase the traffic density in order to simulate risky driver behavior. In the lane changing task for treatment & control model training we incrementally increase the vehicle count from 50 to 150 vehicles accompanied with an equivalent intermittent increase of vehicle density by 100% and having increased randomized risky behavior on episode duration length of 20 seconds. We train unique agents for navigation on each different traffic count environments in our experimentation.
setup corresponding to every vehicle count increase. Also, we plot a comparative analysis performance graph of control and treatment agent across these different environment iterations and calculate ATE of our perturbations. Similarly, for the u-turn environment we uniformly increase our vehicle count from 3 to 12 with incremental increase of vehicle count of 3. Also, for two-way traffic we reduce original environment length to 2/3rd of the original and incrementally increase the vehicle traffic count from base 5 in direction and 2 in opposite direction vehicles to 15 in direction and 6 in opposite direction vehicles. For collision prone treatment in intersection task driving tasks we rewire the randomization behavior to a more risky one with our acceleration and deceleration tuning. We also increase the vehicle count from 10 to 30 incrementally with an interval gap of 5 vehicles and alongside we incrementally increase the spawning probability by 0.1 until it reaches its maximum value. Finally, for roundabout tasks we incrementally increase the traffic from 5 to 15 vehicles with risk prone randomization in our treatment environment for agent training and performance comparison with control baseline. Each of these environment configurations requires the environment to be rebuilt iteratively. Additionally, repeated model training against each new treatment & control environment samples with different initiating random seeds. For evaluating our model’s performance we have kept the traffic as constant in our test population set against the corresponding treatment and control environment on which agents were trained. And we have only changed the risky behavior in treatment and control environment sets to calculate the ATE for measuring causal performance improvement.

We use the DQN reinforcement learning modelling technique with ADAM optimizer for our experiment with a learning rate of 5e-4. Our discount factor used is 0.99 and environment observation vector data is fed in a batch size of 100. Our agents are trained over 3072 episodes until they converge to average reward from a given driving environment. We use the dueling network design which utilizes advantage function $A(s,a)$ which helps in estimating state-action value function $Q(s,a)$ for state-action pairs more precisely. This is done by splitting the network into two streams, value and advantage ones which share some base hidden layers. The shared network consists of 3 fully connected layers of 256, 192 and 128 units respectively. Value and advantage stream consists of 2 layers of 128 units each. The final output of these streams is also fully connected to the network. The value stream has 1 output of the calculated value function for a given state. And the advantage stream has $n_a$ outputs representing the number of discrete possible actions for a given state. The output vectors from these two streams are combined to calculate the state-action value function estimate with the help from below stated equation, refer Figure 6 for model architecture.

$$Q(s,a) = V(s) + A(s,a)$$ (5)

V. RESULTS

Another randomization factor in our experimentation setup involves initial random seed values. This is done for adding random behavior to tasks like randomizing vehicle acceleration & deceleration, spawning vehicles, vehicle location etc. Therefore, we test our trained treatment and control models against the risk-prone and regular driving environments with different randomization seed values to average any anomalous results. Hence, we measure the ATE by evaluating our agents against several randomized seed values for risk prone and regular driving environments. This calculation is simply summarized by the equation below where summation of first two terms evaluate the average reward calculated from treatment models in risk prone and regular environment samples. The last two terms calculate the same for control agent in both these environment samples again respectively and the test set sample count is expressed as $N_T = N_C = 100$. The associated difference of these quantities gives us the ATE for performance improvements in our robust agents trained on perturbed environments as explained earlier in the problem formulation section.

$$ATE = 1/N_T N_C \left[ \sum_{i=1}^{N_T} R_T^{avg} + \sum_{i=1}^{N_C} R_C^{avg} - \sum_{i=1}^{N_T} R_T^{avg} + \sum_{i=1}^{N_C} R_C^{avg} \right]$$ (6)

ATE results from Figure 7 clearly demonstrates to us the advantage of teaching agents these risk prone critical scenarios. Across all tasks we observe that as traffic density increases like real-life scenarios in heavily populated cities the positive effect of knowing perturbed scenarios is pronounced for our robust treatment agents. More importantly there is also

1Baseline DQN agent implementations referenced from: github.com/eleurent/rl-agents
2For better readability purposes we converted the calculated ATE values to their respective percentages in Figure 7.
Fig. 7. ATE plots for quantifying the performance improvements attained by leveraging the knowledge learnt from collision prone driving scenarios.

VI. CONCLUSION

Our experiments from this paper provide insights into the importance of using deliberate interventions of collision prone behaviors while training agents for stochastic processes like autonomous driving. By using the MDP formulation of the discussed driving scenarios with collision simulation perturbations we were able to generate more robust agents. Our treatment model experimentation setup used episode data from traffic clogged lanes and risky randomized behavior during training which finally resulted in positive ATE result values. Which proved that agents trained with a wider range of collision prone scenarios perform better than the existing vanilla simulation agents. Also, we casually quantified the impact of our interventions for the discussed model free learning DQN technique which assisted us in accurately estimating the performance improvements. For every driving scenario environment our new agents produced better results and were proved to be better collision deterrents. Therefore, underscoring the importance of learning valuable lessons from risk prone scenario simulations for creating safe autonomous driving agents.

REFERENCES

[1] D. Elliott, W. Keen, and L. Miao, “Recent advances in connected and automated vehicles,” journal of traffic and transportation engineering (English edition), vol. 6, no. 2, pp. 109–131, 2019.
[2] S. A. Bagloee, M. Tavana, M. Asadi, and T. Oliver, “Autonomous vehicles: challenges, opportunities, and future implications for transportation policies,” Journal of modern transportation, vol. 24, no. 4, pp. 284–303, 2016.
[3] J. Joy and M. Gerla, “Internet of vehicles and autonomous connected car-privacy and security issues,” in 2017 26th International Conference on Computer Communication and Networks (ICCCN). IEEE, 2017, pp. 1–9.
