Dampen the Stop-and-Go Traffic with Connected and Automated Vehicles – A Deep Reinforcement Learning Approach*

Liming Jiang[1], Yuanchang Xie[1], Xiao Wen[1], Danjue Chen[1], Tienan Li[1], Nicholas G. Evans[2]

Abstract— **Stop-and-go traffic poses significant challenges to the efficiency and safety of traffic operations, and its impacts and working mechanism have attracted much attention. Recent studies have shown that Connected and Automated Vehicles (CAVs) with carefully designed longitudinal control have the potential to dampen the stop-and-go wave based on simulated vehicle trajectories. In this study, Deep Reinforcement Learning (DRL) is adopted to control the longitudinal behavior of CAVs and real-world vehicle trajectory data is utilized to train the DRL controller. It considers a Human-Driven (HD) vehicle tailed by a CAV, which are then followed by a platoon of HD vehicles. Such an experimental design is to test how the CAV can help to dampen the stop-and-go wave generated by the lead HD vehicle and contribute to smoothing the following HD vehicles' speed profiles. The DRL control is trained using real-world vehicle trajectories, and eventually evaluated using SUMO simulation. The results show that the DRL control decreases the speed oscillation of the CAV by 54% and 8%-28% for those following HD vehicles. Significant fuel consumption savings are also observed. Additionally, the results suggest that CAVs may act as a traffic stabilizer if they choose to behave slightly altruistically.**

I. INTRODUCTION

The stop-and-go traffic shockwaves is an interesting and important phenomenon [1]. Small perturbations in a lead vehicle's speed profile could be amplified as they are passed on to following vehicles and this creates stop-and-go waves broadcast backwards (i.e., traveling upstream), which results in wasted fuel consumption, additional traffic emissions, increased likelihood of rear-end crashes, and congestion. It is concluded [2] that shorter reaction time and better sharing of vehicle maneuver information are among the keys to address the stop-and-go traffic issue. Therefore, CAV appears to be an ideal candidate solution and has attracted much attention recently. However, there are still several open problems for CAV control that are worth exploring, such as cooperative merging [3], [4], and mitigating the propagation of stop-and-go shockwaves in traffic [5]–[7]. In this study, we focus on the last challenge: minimizing the traffic oscillation in a platoon of vehicles.

Previous studies on this subject mostly [5]–[7] use formula-based approaches to control the behavior of CAV with the goal to dampen the stop-and-go traffic. This research adopts a Deep Reinforcement Learning (DRL) approach to see if CAVs can learn optimal control strategies through interacting with human drivers. The highlights of this study are summarized as follows: (1) Instead of using a closed-loop ring network and assuming the location and maneuver information of all vehicles is known as in previous studies [5], [8], we consider a long straight road segment and take only the CAV and its lead vehicle's state as the algorithm input; (2) To test if our DRL model can work in real word, the speed profile of the lead HD vehicle is sampled from field collected vehicle trajectories in naturalistic driving settings; (3) The proposed DRL control only takes a few input parameters based on the state of the CAV and its lead vehicle. CACC is adopted as a baseline for comparison; and (4) Different reward functions are analyzed and compared.

II. BACKGROUND

A. Related work

Instead of using traditional control theory and formula-based analytic solutions to optimize the behavior of AVs with various objectives, some studies [9]–[13] have tried to adopt machine learning algorithms, especially reinforcement learning. By carefully choosing the state representation and reward function, a RL agent (i.e., CAV) is able to learn how to best regulate its longitudinal behavior based on incentives (rewards) received during interactions with other vehicles. The first work using RL to control AV in a connected environment was conducted by Desjardins and Chaib-draa [12] They concluded that RL-based control could be a promising approach to ensure a safe longitudinal following behavior of CAV to its front vehicle. After that, RL has been widely adopted in CAV behavior modeling, such as longitudinal control [13] and merge control [4], [11].

Two most relevant studies to this paper were conducted by Wu [8] and Qu [14]. They both considered a closed-loop ring road network, which was loaded with some Human Driven (HD) vehicles and one CAV controlled by the RL algorithm. The RL controlled CAV was assumed to have a global (or complete) view of the environment (i.e., speeds and positions of all vehicles), and the CAV learned to address the stop-and-go traffic by maximizing its reward function, which was defined based on vehicles' speed and headway. Although the concept and results of both studies are interesting, assuming a global view is restrictive. Also, their RL controllers were trained completely based on simulated data and a ring network, which may not accurately reflect vehicle maneuvers in practice.

Some other relevant RL studies focused on Cooperative Adaptive Cruise Control (CACC). They adopted RL to train CAVs so that CAVs can stably follow the lead vehicle.

* Research supported by U.S. National Science Foundation (Award #1734521 and #1826162).

[1] Department of Civil and Environmental Engineering, University of Massachusetts (UMass) Lowell, Lowell, MA 01854 USA

[2] Department of Philosophy, UMass Lowell, Lowell, MA 01854 USA

Corresponding Author: Yuanchang Xie, yuanchang_xie@uml.edu
Emails: liming_jiang@student.uml.edu, xiao_wen@student.uml.edu, danjue_chen@uml.edu, tienan_li@student.uml.edu, nicholas_evans@uml.edu

978-1-7281-8995-6/21/$31.00 ©2021 IEEE

However, these studies did not consider using the RL-controlled CAVs to dampen the impacts of stop-and-go shockwave on vehicles following them (i.e., CAVs). In other words, these CAVs behave selfishly without considering vehicles behind them.

CAVs can be trained/designed to behave selfishly or a little altruistically. An interesting but not fully understood question is how and to what extent these different behaviors may affect traffic operations, which is also the motivation of this study. To our best knowledge, this study is the first attempt to use Deep Reinforcement Learning (DRL) approach for CAV control that builds altruism into the control objective (e.g., dampening stop-and-go wave) and also consider real-world vehicle trajectories instead of simulated ones for training.

B. Deep Deterministic Policy Gradient (DDPG)

With Reinforcement Learning (RL) control, each CAV is treated as a RL agent. The agent learns optimal control policies through its interactions with the environment (i.e., surrounding vehicles). Good control policies are rewarded while bad ones are penalized. Over time the agent learns to adjust its behavior to maximize the long-term reward or return.

In this research, our goal is to optimally adjust the CAV's acceleration. Given that the action space is continuous, the Deep Deterministic Policy Gradient (DDPG) algorithm proposed by Google [15] is adopted. DDPG is an actor-critic and model-free RL algorithm. A brief introduction of DDPG is provided below, and more detailed information can be found in the original paper.

Actor-critic RL methods combine the advantages offered by both value-based and policy-based methods by employing an actor (to execute an action) and a critic (to evaluate the action from the actor). This allows actor-critic RL to be more sample efficient [15].

After a minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) is sampled from the replay buffer. The target y_i is calculated as

$$y_i = r_i + \gamma Q^i(s_{i+1}, \mu^{'} \left(s_{i+1} \middle| \theta^{\mu^{'}} \right) \theta^{Q^{'}}) \tag{1}$$

Then the model of critic is updated by minimizing the loss:

$$L = \frac{1}{N} \sum_i \left(y_i - Q(s_i, a_i | \theta^Q)\right)^2 \tag{2}$$

Accordingly, the sampled policy gradient is adopted to update the actor policy.

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_{\mathrm{a}} Q(s, a | \theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i} \tag{3}$$

III. METHODOLOGY

A. Simulation Environment

The scenario considered in this study is a 70-kilometer single-lane road segment created using SUMO (Simulation of Urban MObility). During the simulation, a 10-vehicle platoon is created as shown in Figure 1. The first vehicle is a Human Driven (HD) vehicle that creates the stop-and-go traffic pattern, which is followed by a CAV and eight other HD vehicles. It is assumed that the lead vehicle shares its information with the CAV via V2V communication in real time. The CAV is controlled by a RL agent with the purpose to dampen the stop-and-go shockwave. All other following HD vehicles' behaviors are governed by a modified version of Krauss model [16].

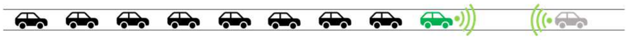

Figure 1. simulation scenario (1^{st}vehicle: stop-and-go pattern; 2^{nd} vehicle: CAV controlled by DRL; 3^{rd}- 10^{th}vehicle: human drivers)

Our goal is to ensure that the trained RL agent can tackle realistic car-following tasks. Therefore, the speed profile of the lead vehicle is sampled from the trajectories of a field observed *highD* dataset [17] and reflects real-world driving behavior in congested traffic scenarios.

All *highD* trajectories were captured on freeways and most of the vehicles were traveling in free-flow mode. Since free-flow traffic does not tell us much about CAV's capability of absorbing the speed oscillation of the lead vehicle, a congested road segment in AM peak is hand-picked and only trajectories with a maximum speed less than 18 m/s and a standard deviation of speed more than 2 m/s are selected. These trajectories are concatenated to generate the speed profile of the lead vehicle in this study.

To concatenate the sampled *highD* trajectories, a gentle acceleration rate ($\pm 0.2 m/s^2$) is adopted to stitch these trajectories together. For instance, if the end speed of the previous trajectory is smaller than the start speed of the next trajectory, a positive acceleration rate of $0.2 m/s^2$ is adopted to fill the speed gap of the two-consecutive trajectories. Otherwise, a deceleration rate of $-0.2 m/s^2$ would be adopted.

B. Parameter Modeling

In RL, three critical factors are: system state representation, action space definition, and reward function. In this section, the three parameters are described in detail.

- State Representation

The system state represents what an agent can actually senses regarding the surrounding environment and itself. In this study, we consider the maneuver information of the lead vehicle and the ego vehicle (the RL-controlled CAV). It is assumed that they are in a connected environment and there are on-board devices that let these two vehicles share movement dynamics information in real time. Specifically, the state representation is defined as:

$$\{\Delta s, v_{lead}, a_{lead}, v_{ego}, a_{ego}\}$$

where, Δs is the distance between the lead vehicle and the ego vehicle. v_{lead} and $a_{lead}$are the speed and acceleration of the lead vehicle, respectively. Accordingly, v_{ego} and α_{ego} are the speed of acceleration of the ego vehicle, respectively.

- Action Space

Action space defines the range of actions the RL agent can execute at each time step. To reflect the realistic characteristics of regular vehicles. we restrict the acceleration rate (i.e., actions pace) to be in the range of (-3, 2) in m/s^2.

- Reward Function

Reward function is a mapping from state and action representations to rewards received by agent at each time step. Reward function serves as motivations to agent and is the key design feature one can control to regulate agent's behavior. A good design of reward function helps agent learn the intended behavior and facilitates the learning process to converge to an optimum relatively fast.

Our reward function design consists of four main goals. The first goal is for ensuring safety. Depending on the time headway between the lead and ego vehicles, the agent will be given the safety reward defined in Eq. (1).

$$Reward_{headway} = \begin{cases} -100, if\ headway \leq 0 \\ -100 + \sqrt{100^2(1-(x-1)^2)}, \\ if\ 0 < headway \leq 1 \\ 0, if\ headway > 1 \end{cases} \quad (1)$$

It is obvious that one can design a safe CAV control to dampen the stop-and-go waves by letting the ego vehicle (i.e., the CAV) travel at a constant but very low speed. As long as the ego vehicle's speed is much less than the lead vehicle's average speed, the ego vehicle most likely will not need to decelerate and can maintain a safe headway with the lead vehicle. However, this approach would constantly increase the ego vehicle's gap to the lead vehicle, thus making it a moving bottleneck and increasing the anxiety of drivers behind it. Therefore, another goal is considered to ensure that the ego vehicle maintains a safe headway but also a reasonably fast speed.

In our design, the speed goal is broken down into two parts as in Eq. (2) , in which $Reward_{speed}$ is the reward regarding the speed of the ego vehicle, v_{Ept} is the expected speed or speed limit, and v_{ego} is the speed of the ego vehicle. The first part of Eq. (2) encourages the ego vehicle to maintain a high speed, while the second part penalizes the ego vehicle for going slower than the expected speed but does not reward it for going faster than it. By combining the two parts, the goal is for the ego vehicle to catch up with the lead vehicle but not to travel faster than it.

$$Reward_{speed} = v_{Ept} - Max(, v_{Ept} - v_{ego}) \quad (2)$$

The third goal is to dampen the speed variation. In other words, when the lead vehicle executes a hard deceleration, the ego vehicle should be able to predict that and take proactive actions (e.g., maintain a large time headway in anticipation of the hard deceleration) so that a hard deceleration is not needed for the ego vehicle and the following HD vehicles would not need to brake hard either.

To achieve this goal, a speed penalty term defined in Eq. (3) is considered if the headway h is smaller than a critical headway value h_c. The rationale behind this is that the ego vehicle is not supposed to travel faster than its lead vehicle when it is approaching the lead vehicle. If it does (e.g., in the situation that the lead vehicle decelerates), the ego vehicle should be rewarded by reducing its speed relative to the lead vehicle.

$$Reward_{speeddiff} = (v_{ego} - v_{lead}) * (h - h_c) \quad (3)$$

Where $v_{ego} - v_{lead}$ is speed difference between the ego vehicle and the lead vehicle, h is the current time headway of the ego vehicle, and h_c is the critical headway set to be 1 s in this study. This reward is calculated only if the speed of ego vehicle v_{ego} is less than lead vehicle's speed v_{lead} and headway of ego vehicle h is less than the critical headway h_c. Based on this equation, more penalty is given to the DRL agent when it drives faster than its lead vehicle and keep a shorter than critical gap to its lead vehicle.

Another reward function term for achieving the third goal is to penalize large acceleration rates. Large acceleration rates (either negative or positive) should be penalized to ensure smooth driving and help to reduce speed oscillation. This term is defined as follows:

$$Reward_{acc} = -a_{ego}^2 \quad (4)$$

As in Eq. (4), to further penalize large accelerations the acceleration of ego vehicle is squared. The following Eq. (5) is the complete reward function that includes all previously discussed reward terms.

$$Reward = \alpha * Reward_{headway} + \beta * Reward_{speed} + \gamma * Reward_{speeddiff} + \delta * Reward_{acc} \quad (5)$$

In Eq. (5) α, β, γ and δ are hyperparameters of this reward function, which specify the weights for each reward terms. After careful hyperparameter tuning, we decide to choose the following values: $\alpha = 1, \beta = 1, \gamma = 1$, and δ = 4.

IV. RESULTS

In this section, key results are presented and compared to a baseline model (with all HD vehicles) both qualitatively and quantitatively. In addition, another baseline scenario is modeled which replaces the DRL-controlled CAV by a vehicle controlled by Cooperative Adaptive Cruise Control (CACC) [18].

A. Evaluation Methodology

The trained CAV DRL control is coded and evaluated in SUMO simulation. The HD and CACC scenarios serve as the baselines to demonstrate the superiority of the proposed DRL model.

The following rolling mean and standard deviation of speed are adopted to measure the capability of the DRL-controlled CAV to absorb hard decelerations of the lead vehicle.

$$\bar{x}_k^T = \frac{1}{T}\sum_{i=k}^{k+T} x_i \quad (4)$$

$$s_k^T = \sqrt{\frac{1}{T-1}\sum_{i=k}^{k+T}(x_i - \bar{x}_k^T)^2} \quad (5)$$

Where T is the rolling time window length, $\bar{x}_k^T$ is the average speed starting at the k^{th} time step of length T, and s_k^T is the rolling standard deviation of speeds starting at the k^{th} time step. To measure the overall speed variation over the entire evaluation period, the average of all rolling standard deviations s_i^T is utilized as defined below.

$$\bar{s}^T = \frac{1}{end - start - T - 1}\sum_{i=start}^{end-T} s_i^T \quad (6)$$

B. *Experiment Results*

The total evaluation simulation run time is about two hours. Simulated vehicle trajectories from a randomly selected 2.5-minute time window are extracted (See Figure 2) to compare the performances of the proposed model and the baseline. Based on the trajectories, vehicle speed and acceleration profiles are also plotted and presented in Figures 3 and 4.

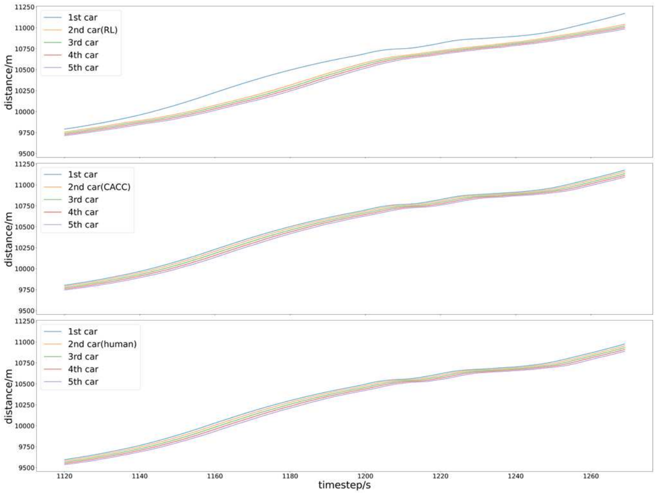

Figure 2. Trajectory of First 5 Vehicles (top: RL agent as 2nd car; mid: CACC as 2nd car; bottom: human driver as 2nd car)

As can be seen in the trajectory figure (Figure 3), the DRL-controlled CAV tends to keep a longer distance (about 50 meters on average with a standard deviation of 30 meters) to its front car than human drivers. Although this distance is not explicitly defined by the reward function in our design, the DRL agent figures out by itself that in order to avoid hard and/or frequent accelerations/decelerations, it needs to increase the gap to its lead vehicle and use that gap as a buffer zone to absorb the stop-and-go shockwave.

Another interesting finding from Figure 3 is that the HD follower of the lead vehicle tends to copy and exaggerate the lead vehicle's behavior. For example, every time the lead vehicle decelerates to a speed v, the following HD vehicle tends to decelerate to an even smaller speed compared to v. The CACC vehicle performs similar to the HD vehicle, except that its behavior is a little more stable. The proposed DRL control performs the best. It tends to dampen the speed oscillation of the lead vehicle and takes a less extreme action compared to its lead vehicle. The acceleration profiles in Figure 4 also show that DRL can keep the acceleration oscillation within a much smaller range compared to CACC and HD vehicles.

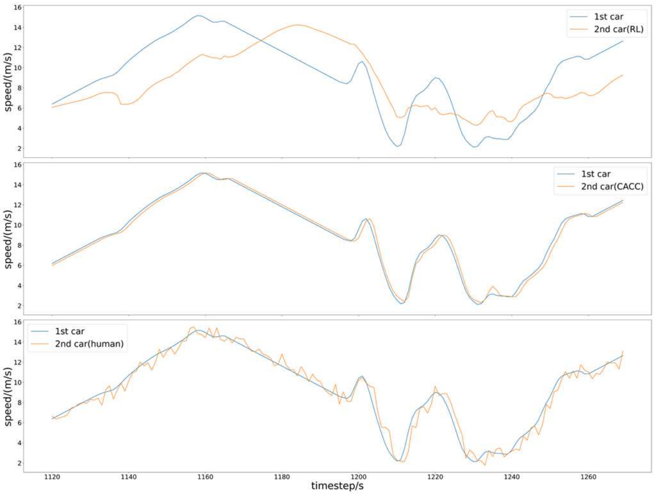

Figure 3. Speed Profile of First Two Cars in Fleets (top: RL agent as 2nd car; mid: CACC as 2nd car; bottom: human driver as 2nd car)

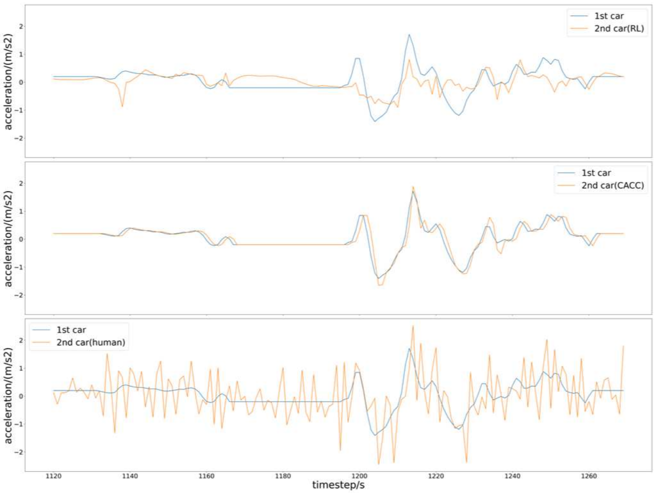

Figure 4. Acceleration Profile of First Two Cars in Fleets (top: RL agent as 2nd car; mid: CACC as 2nd car; bottom: human driver as 2nd car)

Given the above comparison results, a valid question is that is the DRL control going to increase the overall travel time. For example, the CAV travels at a very low and constant speed. Although this can lead to a smooth trajectory, it will take the CAV much long time to travel the same distance than a HD vehicle with a stop-and-go trajectory. Based on the simulated results, the DRL-controlled CAV is able to not only absorb the stop-and-go shockwave created by the lead vehicle, but also travel at the same average journey speed as a HD vehicle. Note that all the CAV, CACC and HD vehicle speeds are constrained by the lead vehicle.

To quantify the effects of how stop-and-go waves get dampened with CAV and HD vehicle, the average rolling speed standard deviations for vehicles at different positions are calculated and presented in Table 1.

TABLE I. $\bar{s}^T$ FOR CAV, CACC, AND HD VEHICLES

Vehicle Position in the Platoon	**Performance Measure**		
	$\bar{s}^T$ **(CAV - RL agent)**	$\bar{s}^T$ **(CACC)**	$\bar{s}^T$ **(all human)**
1st	0.33	0.33	0.33
2nd	0.21 (-64%)	0.34 (-41%)	0.58
3rd	0.56 (-26%)	0.63 (-17%)	0.76
4th	0.71 (-14%)	0.77 (-7%)	0.83
5th	0.81 (-11%)	0.86 (-5%)	0.91
6th	0.86 (-9%)	0.90 (-5%)	0.95
7th	0.91 (-9%)	0.94 (-6%)	1.00
8th	0.93 (-12%)	0.97 (-8%)	1.06
9th	0.96 (-15%)	1.00 (-11%)	1.13
10th	0.98 (-12%)	1.00 (-10%)	1.11

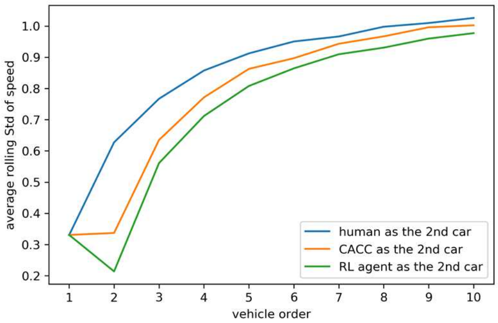

Figure 5. Speed Oscillations of Different Vehicles

In Table 1, the first vehicles in CAV (i.e., RL agent), CACC and HD (i.e., all human drivers) scenarios all have the same speed oscillation as their trajectories are sampled from the *highD* dataset and are identical. For the 2nd vehicle in the platoon, the CAV is able to absorb the speed oscillation by 64% compared to its HD vehicle counterpart. While the benefit of CACC is 41% compared to HD vehicle. Interestingly, the CACC seems to copy the behavior of the lead vehicle (see Figure 2 and Figure 3) thus the $\bar{s}^T$ stays about the same. The effect on the 3rd vehicle in the CAV scenario is less significant because the 3rd vehicle is controlled by a human driver. Nevertheless, it still has a 26% reduction in speed oscillation compared to the HD scenario. For the remaining vehicles in the platoon, the speed oscillation reduction benefits are in the range of 9%~15%. In sum, the oscillation reduction benefits reach the peak (over 50% reduction) for the 2nd vehicle in the platoon, and drop to the somewhere near 8% as the stop-and-go wave propagates to the 7th vehicle, and increase slightly afterwards. The same trend can be found for CACC although the magnitude of oscillation reductions is less than CAV.

The fuel consumption reduction benefits are also studied. The emission model HBEFA [19] is used to quantify the fuel consumption for each vehicle in both scenarios, and the comparison results are plotted in Figure 6. The overall patterns for fuel consumption savings and speed oscillation reductions are similar. Again, the CAV (RL agent) is able to save more fuel than CACC for a fleet.

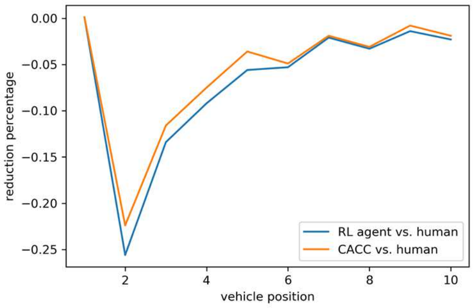

Figure 6. Fuel Consumption of CAV Scenario Compared to HD Scenario.

C. Other Reward Functions

As mentioned in the introduction section, several other researchers also investigated this problem. However, they considered different state representations, reward functions, and simulation scenarios. Wu [8] used a reward function based on speed which encourages vehicles on a ring road to travel as close to speed limit as possible. Qu [14] adopted this idea and further added a time headway term to address safety concerns. They both trained their RL agents on a ring road where the first vehicle could catch the last vehicle in the platoon, and assumed the RL agents have a global view of the environment (i.e., speeds and positions of all vehicles). As a comparison, we adopted their reward function definitions and trained the RL agent on a straight single-lane road to see if their reward functions will properly guide the CAV.

The agents trained according to Wu's reward function (Reward W for short) and Qu's reward function (Reward Q for short) both reach return plateau after 3 epochs (two hours of simulated driving for each epoch). Based on direct observation and trajectory analysis, CAV trained by Reward W drives very cautiously/slowly and the gap between the CAV and the lead vehicle keeps increasing. This is probably due to the non-continuity in the reward function, which still rewards an agent even though it is very close to the lead vehicle and suddenly generates a very large penalty if the agent gets slightly closer (The agent touches the rear end of the lead vehicle). Although Reward W works well on a ring road to stabilize traffic, that success cannot be successfully transferred to calm stop-and-go traffic on a straight road segment.

With Reward Q, the CAV is able to follow the lead vehicle relatively well without creating a large gap. We think this is because of the additional headway term in the reward function, which brings continuity into the reward function and avoids a sudden transition from reward to penalty. However, Reward Q still leads to unstable speed and acceleration profiles as in Figure 7. Under this reward function, the CAV frequently oscillates between the maximum-allowed acceleration and deceleration. The $\bar{s}^T$ as a result of Reward Q reaches 1.67, which is much higher than the HD baseline (0.58). This suggest that Reward Q would significantly amplify the stop-and-go shockwave rather than dampen it.

Overall, the above comparison results suggest that reward functions are critical in training RL CAV control agents. One contribution of this study is that it includes two reward terms specifically designed for dampening the stop-and-go shockwave, and evaluates their effectiveness on a relatively realistic road network.

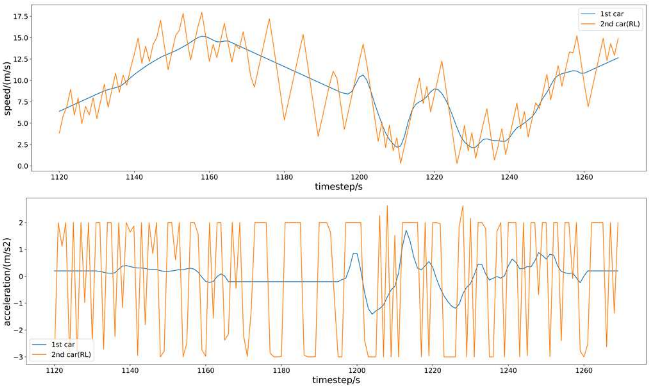

Figure 7. Speed (Top) and Acceleration (Bottom) Profile of First Two Vehicles by Reward Q

V. Conclusion

This study shows that DRL is a promising approach to dampen stop-and-go traffic and generate significant safety and environmental benefits in terms of speed variation and fuel consumption reductions, respectively. These benefits are not only for the ego CAV vehicle, but also for other human driven vehicles following it. This brings up an interesting question for future research: should CAVs behave in its own interest only or altruistically? Also, more work can be done considering multiple DRL-controlled CAVs. Additionally, it would be interesting to model the performance of the DRL model in a multi-lane environment.

References

[1] Y. Sugiyama *et al.*, “Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam,” *New J. Phys.*, vol. 10, no. 3, p. 033001, Mar. 2008, doi: 10.1088/1367-2630/10/3/033001.

[2] X. Li, J. Cui, S. An, and M. Parsafard, “Stop-and-go traffic analysis: Theoretical properties, environmental impacts and oscillation mitigation,” *Transportation Research Part B: Methodological*, vol. 70, pp. 319–339, Dec. 2014, doi: 10.1016/j.trb.2014.09.014.

[3] T. Ren, Y. Xie, and L. Jiang, “New England merge: a novel cooperative merge control method for improving highway work zone mobility and safety,” *Journal of Intelligent Transportation Systems*, vol. 0, no. 0, pp. 1–15, Sep. 2020, doi: 10.1080/15472450.2020.1822747.

[4] T. Ren, Y. Xie, and L. Jiang, “Cooperative Highway Work Zone Merge Control Based on Reinforcement Learning in a Connected and Automated Environment,” *Transportation Research Record*, 2020.

[5] R. E. Stern *et al.*, “Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments,” *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 205–221, Apr. 2018, doi: 10.1016/j.trc.2018.02.005.

[6] J. I. Ge and G. Orosz, “Dynamics of connected vehicle systems with delayed acceleration feedback,” *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 46–64, Sep. 2014, doi: 10.1016/j.trc.2014.04.014.

[7] C. Wu, A. M. Bayen, and A. Mehta, “Stabilizing Traffic with Autonomous Vehicles,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, May 2018, pp. 1–7, doi: 10.1109/ICRA.2018.8460567.

[8] Wu, Cathy, “Learning and Optimization for Mixed Autonomy Systems-A Mobility Context.” UC Berkeley, 2018.

[9] A. Khodayari, A. Ghaffari, R. Kazemi, and R. Braunstingl, “A Modified Car-Following Model Based on a Neural Network Model of the Human Driver Effects,” *IEEE Trans. Syst., Man, Cybern. A*, vol. 42, no. 6, pp. 1440–1449, Nov. 2012, doi: 10.1109/TSMCA.2012.2192262.

[10] J. Morton, T. A. Wheeler, and M. J. Kochenderfer, “Analysis of Recurrent Neural Networks for Probabilistic Modeling of Driver Behavior,” *IEEE Trans. Intell. Transport. Syst.*, vol. 18, no. 5, pp. 1289–1298, May 2017, doi: 10.1109/TITS.2016.2603007.

[11] P. Wang and C.-Y. Chan, “Formulation of deep reinforcement learning architecture toward autonomous driving for on-ramp merge,” in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Oct. 2017, pp. 1–6, doi: 10.1109/ITSC.2017.8317735.

[12] C. Desjardins and B. Chaib-draa, “Cooperative Adaptive Cruise Control: A Reinforcement Learning Approach,” *IEEE Trans. Intell. Transport. Syst.*, vol. 12, no. 4, pp. 1248–1260, Dec. 2011, doi: 10.1109/TITS.2011.2157145.

[13] M. Zhu, Y. Wang, Z. Pu, J. Hu, X. Wang, and R. Ke, “Safe, Efficient, and Comfortable Velocity Control based on Reinforcement Learning for Autonomous Driving,” *arXiv:1902.00089 [cs, stat]*, Oct. 2019, Accessed: Feb. 13, 2020. [Online]. Available: http://arxiv.org/abs/1902.00089.

[14] X. Qu, Y. Yu, M. Zhou, C.-T. Lin, and X. Wang, “Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: A reinforcement learning based approach,” *Applied Energy*, vol. 257, p. 114030, Jan. 2020, doi: 10.1016/j.apenergy.2019.114030.

[15] T. P. Lillicrap *et al.*, “Continuous control with deep reinforcement learning,” *arXiv:1509.02971 [cs, stat]*, Jul. 2019, Accessed: Mar. 07, 2020. [Online]. Available: http://arxiv.org/abs/1509.02971.

[16] S. Krauss, P. Wagner, and C. Gawron, “Metastable states in a microscopic model of traffic flow,” *Phys. Rev. E*, vol. 55, no. 5, pp. 5597–5602, May 1997, doi: 10.1103/PhysRevE.55.5597.

[17] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, “The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems,” in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, HI, Nov. 2018, pp. 2118–2125, doi: 10.1109/ITSC.2018.8569552.

[18] B. van Arem, C. J. G. van Driel, and R. Visser, “The Impact of Cooperative Adaptive Cruise Control on Traffic-Flow Characteristics,” *IEEE Trans. Intell. Transport. Syst.*, vol. 7, no. 4, pp. 429–436, Dec. 2006, doi: 10.1109/TITS.2006.884615.

[19] P. D. Haan and M. Keller, “Modelling fuel consumption and pollutant emissions based on real-world driving patterns: the HBEFA approach,” *IJEP*, vol. 22, no. 3, p. 240, 2004, doi: 10.1504/IJEP.2004.005538.