



Tactical decision making for autonomous trucks by deep reinforcement learning with total cost of operation based reward

Downloaded from: <https://research.chalmers.se>, 2026-05-12 16:13 UTC

Citation for the original published paper (version of record):

Pathare, D., Laine, L., Haghiri Chehreghani, M. (2026). Tactical decision making for autonomous trucks by deep reinforcement learning with total cost of operation based reward. *Artificial Intelligence Review*, 59(1).
<http://dx.doi.org/10.1007/s10462-025-11448-8>

N.B. When citing this work, cite the original published paper.



Tactical decision making for autonomous trucks by deep reinforcement learning with total cost of operation based reward

Deepthi Pathare^{1,3} · Leo Laine^{2,3} · Morteza Haghiri Chehreghani¹

Received: 16 April 2024 / Accepted: 4 November 2025
© The Author(s) 2025

Abstract

We develop a *deep reinforcement learning* framework for tactical decision making in an autonomous truck, specifically for Adaptive Cruise Control (ACC) and lane change maneuvers in a highway scenario. Our results demonstrate that it is beneficial to separate high-level decision-making processes and low-level control actions between the reinforcement learning agent and the low-level controllers based on physical models. In the following, we study optimizing the performance with a realistic and multi-objective reward function based on Total Cost of Operation (TCOP) of the truck using different approaches; by adding weights to reward components, by normalizing the reward components and by using curriculum learning techniques.

Keywords Autonomous trucks · Deep reinforcement learning · Curriculum learning · Total cost of operation

1 Introduction

The efficiency and safety of transport networks have a huge impact on the socio-economic development of the globe. A significant part of this network is freight transport, of which more than 70% is carried out by trucks (De Jong et al. 2004). The modeling of these com-

✉ Deepthi Pathare
pathare@chalmers.se
Leo Laine
leo.laine@chalmers.se
Morteza Haghiri Chehreghani
morteza.chehreghani@chalmers.se

¹ Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Göteborg 41296, Sweden

² Department of Mechanics and Maritime Sciences, Chalmers University of Technology, Göteborg 41296, Sweden

³ Safe and Efficient Driving, Volvo Group of Trucks Technology, Göteborg 41715, Sweden

plex networks, with a particular focus on traffic scenarios, including trucks and their Total Cost of Operation (TCOP) is thus of paramount importance to developing sustainable traffic solutions.

At a mesoscopic scale, a truck can significantly affect the surrounding traffic (Moreno et al. 2018), primarily due to its comparatively larger size and length (Yang et al. 2015). It also needs more cooperation from surrounding vehicles in order to perform specific maneuvers, such as lane change in multiple-lane dense traffic scenarios (Nilsson et al. 2018). Further, if it is a Long Combination Vehicle (LCV), its influence on the safety and evolution of the surrounding traffic is substantial (Grislis 2010).

Modern vehicles, including trucks, are equipped with a number of features that enhance their performance at different levels. For example, modern trucks operate in connected networks, constantly exchanging data related to their location and performance with their clients. They are also equipped with Advanced Driver Assistance Systems (ADAS) to assist the driver with complex driving tasks and to enhance safety (Shaout et al. 2011; Jiménez et al. 2016). Adaptive Cruise Control (ACC) is such a driver assistance function that provides longitudinal control and maintains a safe distance with the vehicle ahead (Xiao and Gao 2010).

Artificial Intelligence (AI) and Machine Learning have revolutionized the connectivity and autonomy of vehicles, including trucks. With the integration of sensors, cameras, and sophisticated onboard systems, machine learning algorithms can analyze vast amounts of real-time data, allowing trucks to make intelligent decisions on the road. Continuous advancements in machine learning continue to push the boundaries of connectivity and autonomy, transforming the trucking industry towards a more efficient and intelligent future.

A widely used machine learning framework in the context of autonomous systems is Reinforcement Learning (RL). RL overcomes the challenges of traditional search based methods such as A* search, which lack the ability to generalize to unknown situations in a non-deterministic environment and are computationally expensive (Sutton and Barto 2018). RL methods also have significant advantages over traditional car-following and lane change models which are based on predefined rules and assumptions. Behavior of traditional models are deterministic and do not adapt beyond their programmed logic. On the otherhand, advanced RL methods such as Deep Reinforcement Learning (DRL) can adapt to non-deterministic environments with varying surrounding vehicle behaviors and make informed decisions in complex scenarios by processing high-dimensional state representations. RL has been increasingly used to solve complex problems related to autonomous car driving, such as navigation, trajectory planning, collision avoidance, and behavior planning (Sallab et al. 2017; Shalev-Shwartz et al. 2016; Kiran et al. 2021). The application of this framework to autonomous trucks is a relatively new area of research. An important contribution in this direction is the work (Hoel et al. 2020), which implements an RL framework for autonomous truck driving in SUMO. They study how a Bayesian RL technique, based on an ensemble of neural networks with additional randomized prior functions (RPF), can be used to estimate the uncertainty of decisions in autonomous driving. On the other hand, more results exist when we consider autonomous driving of passenger cars. The study in (Desjardins and Chaib-draa 2011) develops a cooperative adaptive cruise control (CACC) using RL for securing longitudinal following of a front vehicle using vehicle-to-vehicle communication. The CACC system has a coordination layer that is responsible for the selection of high-level actions, e.g., lane changing or secure vehicle following. The action layer,

which is a policy-gradient RL agent, must achieve this action by selecting the appropriate low-level actions that correspond to the vehicle's steering, brakes, and throttle. Another study in Zhao et al. (2013) proposes a Supervised Actor-Critic (SAC) (Konda and Tsitsiklis 1999; Haarnoja et al. 2018) approach for optimal control of ACC. Their framework contains an upper level controller based on SAC, which chooses desired acceleration based on relative velocity and distance parameters of leading and following vehicles. A low level controller receives the acceleration signal and transfers it to the corresponding brake and/or throttle control action. In Lin et al. (2021), the authors develop a DRL framework for ACC and compare the results with Model Predictive Control. In all these studies, the RL agent performs step control of acceleration or other driving maneuvers such as braking and steering. For this reason, it can take longer time for the agent to achieve maximum speed even though there is no leading vehicle in front, making the process apparently inefficient.

There is also only limited prior work available where the RL agent is trained to learn high level actions such as choosing the safe gap with the leading vehicle and the actual speed control is performed by a low level controller. The work in Das and Won (2021) develops a similar system that aims to achieve simultaneous optimization of traffic efficiency, driving safety, and driving comfort through dynamic adaptation of the inter-vehicle gap. It suggests a dual RL agent approach - the first RL agent is designed to find and adapt the optimal time-to-collision (TTC) threshold based on rich traffic information, including both macroscopic and microscopic traffic data obtained from the surrounding environment and the second RL agent is designed to derive an optimal inter-vehicle gap that maximizes traffic flow, driving safety, and comfort at the same time. In Yang et al. (2020), authors develop a Deep Deterministic Policy Gradient (DDPG) - Proportional Integral Derivative (PID) controller for longitudinal control of vehicle platooning. They use the DDPG algorithm to optimize the K_p , K_d and K_i constants for the PID controller which controls the desired speed. Both these studies mainly focus on longitudinal control of vehicles, and the effects of lane change maneuvers are not studied.

We develop ACC, together with lane change maneuvers for an autonomous truck in a highway scenario. We propose an architecture that combines RL with low-level controllers and separate the high level and low level decision making between them to improve safety and efficiency. The results of this architecture are compared to a baseline architecture that solely relies on RL to perform actions. We evaluate the performance with three different RL algorithms: Deep Q-Network (DQN) (Mnih et al. 2015), Advantage Actor-Critic (A2C) (Mnih et al. 2016) and Proximal Policy Optimization (PPO) (Schulman et al. 2017). We design two reward functions for training the RL agents. The first focuses on safety, while the second is based on Total Cost of Operation (TCOP), covering realistic expenses like energy consumption and human resources. We investigate different training methods to handle the complex and realistic reward function based on TCOP. This approach establishes a strong foundation for future research into improving the economic viability and operational efficiency of autonomous driving systems.

This work is an extension of our previous work (Pathare et al. 2023) wherein we additionally, (i) investigate training of RL agent with the complex TCOP based reward function consisting of realistic costs and revenue values with and without normalization (ii) extend our RL framework by incorporating curriculum learning techniques (Bengio et al. 2009; Narvekar et al. 2020) and compare the results with the non-curriculum learning approach. All our results are obtained using Simulation of Urban Mobility (SUMO) (Lopez et al.

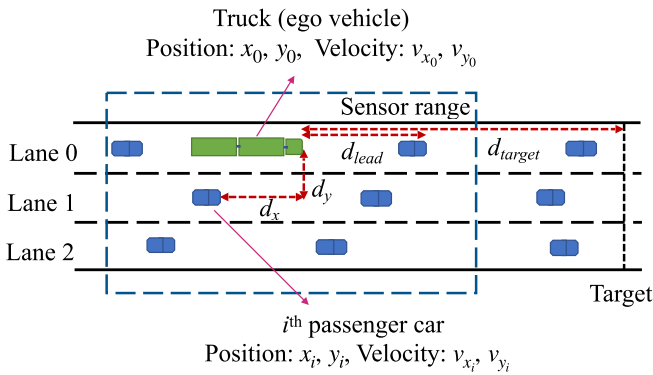


Fig. 1 Schematic diagram of the highway simulation environment. The truck in green color is the ego vehicle

2018), which is a widely used simulation platform for the conceptual development of autonomous vehicles. The tailored RL environment based on SUMO that we developed for heavy vehicle combination driving in highway traffic is provided as open access, which offers a great open source framework for investigating various RL methods in complex settings.¹

2 Tactical decision making with reinforcement learning

2.1 Environment setup

In this work, we consider a dynamic highway traffic environment with an autonomous truck (ego vehicle) and passenger cars as shown in Fig. 1. The maximum speed of the ego vehicle is set to be 25 m/s. Furthermore, 15 passenger cars with speed between 15 m/s and 35 m/s are simulated based on the default Krauss car following model (Krauss et al. 1997) and LC2013 lane change model (Erdmann 2014) in SUMO. The initial position and speed pattern of the surrounding vehicles are set randomly which makes the surrounding traffic nondeterministic throughout the simulation. Starting from the initial position (which will be 800 m after initialization steps in SUMO), the ego vehicle is expected to reach the target set at a distance of 3000 m. This means that the ego vehicle has to drive 2200 m on the highway in each episode. An episode will also terminate if a hazardous situation such as a collision or driving outside the road happens or if a maximum of 500 RL steps are executed. The observation or state of the environment at each step contains information about position, speed, and state of left/right indicators for ego vehicle and vehicles within the sensor range. The sensor range used here is 200 m.

2.2 Reinforcement learning framework

Reinforcement learning is a branch of machine learning where an agent acts in an environment and tries to learn a policy, π , that maximizes a cumulative reward function. The policy

¹Source Code: <https://github.com/deepthi-pathare/Autonomous-truck-sumo-gym-env>

defines which action, a , to take, given an observation or state, s . This action leads to a new state of the environment s' , and returns a reward, r .

2.2.1 MDP formulation

A reinforcement learning problem can be modeled as a Markov Decision Process (MDP), which is defined by the tuple (S, A, T, R, γ) , where S is the state space, A is the action space, P is the transition dynamics, R is the reward model, and γ is a discount factor. The tactical decision making problem for autonomous truck in this study is formulated as MDP as follows:

- **State Space (S):** The state includes observations from ego vehicle and the surrounding vehicles. Following are the observations for the ego vehicle:
 1. Longitudinal speed (v_{x0})
 2. Lane change state ($sign(v_{y0})$)
 3. Lane number
 4. State of left indicator
 5. State of right indicator
 6. Target(leading) vehicle distance (d_{lead})

Following are the observations for each vehicle in the sensor range of the ego vehicle:

1. Relative longitudinal distance from ego vehicle (d_{xi})
 2. Relative lateral distance from ego vehicle (d_{yi})
 3. Relative longitudinal speed with ego vehicle ($v_{xi} - v_{x0}$)
 4. Lane change state ($sign(v_{yi})$)
 5. Lane number
 6. State of left indicator
 7. State of right indicator
- **Action Space (A):** The action space is defined separately for the baseline and new architectures described in section 2.3.
 - **Transition Dynamics(P):** The transition dynamics is defined by SUMO, and is not known to the RL agent.
 - **Reward Function(R):** We design a basic reward function focused on safety aspects to motivate the agent to drive at maximum speed and reach the target without any hazardous situations. This reward function is similar to that in Hoel et al. (2020) except that a reward for reaching the target is added, whereas the penalty for emergency braking is not considered for simplicity. The reward at each time step is given by (1).

$$r(t) = \frac{v_t}{max_v} - I_l P_l - I_c P_c - I_{nc} P_{nc} - I_o P_o + I_{tar} \frac{R_{tar}}{T} \quad (1)$$

Here v_t is the velocity of the vehicle at time step t and max_v is the maximum velocity of the vehicle. I is an indicator function, which takes a value of 1 when the corresponding condition is satisfied and P and R are the corresponding penalty and reward values respectively. The possible conditions are lane change (l), collision (c), near collision (nc), driving outside the road (o) and reaching the target (tar). T is the total time it takes to reach the target. The parameter values used are given in Table 1. Furthermore, we have also designed a complex and multi-objective reward function based on TCOP which is described in section 2.4.

- **Discount Factor**(γ): γ is set to 0.99.

2.2.2 Reinforcement learning algorithms

We choose three reinforcement learning algorithms as listed below for the decision making in baseline and new architectures. The implementations of these algorithms from the stable-baselines3 library with the default hyperparameters are used Raffin et al. (2021).

1. **Deep Q-Network (DQN)** DQN is a reinforcement learning algorithm based on the Q-learning algorithm, which learns the optimal action-value function by iteratively updating estimates of the Q-value for each state-action pair. In DQN, the optimal action-value function in a given environment is approximated using a neural network model. The corresponding objective function is given by,

$$L(\theta) = \mathbb{E}_{s,a,r,s'} \left[\left(y - Q(s, a; \theta) \right)^2 \right] \quad (2)$$

where θ represents the weights of the Q-network, s and a are the state and action at time t , r is the immediate reward, s' is the next state, and $y = r + \gamma \max_{a'} Q(s', a'; \theta^-)$ is the target value. The target value is the sum of the immediate reward and the discounted maximum Q-value of the next state, where γ is the discount factor and θ^- represents the weights of a target network with delayed updates. The objective function is to minimize the mean squared error between the estimated Q-value and the target value, which is then optimized using stochastic gradient descent. Another improvement of DQN over the standard Q learning approaches is the use of a replay buffer and target network. The agent uses replay buffer to store transitions and samples from it randomly during training. This enables the efficient use of past experiences. The target network is a copy of the Q-network with delayed updates. Together with the replay buffer, it improves the learning stability of the model and help prevent over fitting.

2. **Advantage Actor-Critic (A2C)** A2C is a reinforcement learning algorithm that combines the actor-critic method with an estimate of the advantage function. In the actor-critic method, the agent learns two neural networks: an actor network that outputs a

Table 1 Parameter values used in the basic reward function

Parameter	Value
P_l	1
P_c	10
P_{nc}	10
P_o	10
R_{tar}	100

probability distribution over actions, and a critic network that estimates the state-value function. The actor network is trained to maximize the expected reward by adjusting the policy parameters, while the critic network is trained to minimize the difference between the estimated value function and the true value function. Advantage function calculates how better taking an action at a state is compared to the average value of the state which is computed as below:

$$A(s_t, a_t) = r_{t+1} + \gamma V(s_{t+1}) - V(s) \quad (3)$$

Actor uses the computed advantage function value from critic as a feedback and updates the policy parameters θ as,

$$\Delta\theta = \alpha_\theta A(s_t, a_t) \nabla_\theta (\log \pi_\theta(s_t, a_t)) \quad (4)$$

Critic updates its value function parameters w as,

$$\Delta w = \alpha_w A(s_t, a_t) \nabla_w \hat{v}_w(s_t) \quad (5)$$

Here, α_θ and α_w are the learning rates.

3. **Proximal Policy Optimization (PPO)** PPO belongs to the family of policy gradient algorithms and has exhibited great potential in effectively addressing diverse RL problems (Schulman et al. 2017; Svensson et al. 2023). In addition, PPO benefits from simple yet effective implementation and a broad scope of applicability. The PPO algorithm is designed to address the challenges such as the instability that can arise when the policies are rapidly updated. PPO works by optimizing a surrogate objective function that measures the difference between the updated policy and the previous one. The surrogate objective function of PPO is defined as the minimum of two terms: the ratio of the new policy to the old policy multiplied by the advantage function, and the clipped ratio of the new policy to the old policy multiplied by the advantage function. In mathematical terms, this function is given as,

$$L^{CLIP}(\theta) = \hat{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (6)$$

where θ represents the policy parameters, $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the likelihood ratio between the current policy and the previous policy, \hat{A}_t is the estimated advantage of taking action a_t in state s_t and ϵ is a hyperparameter that controls the degree of clipping. The clipping term ensures that the policy update does not result in significant policy changes, which can destabilize the learning process. The PPO algorithm maximizes this objective function using stochastic gradient ascent to update the policy parameters. The expectation is taken over a batch of trajectories sampled from the environment.

2.3 System architectures

2.3.1 Baseline architecture

The baseline architecture we consider in this work is inspired from Hoel et al. (2020) and is illustrated in Fig. 2. Here the decision making is done solely by an RL agent, every 1 s. The action space for the agent is discrete. Each action consists of a combination of longitudinal action and lateral action. Longitudinal actions are speed changes of 0, 1, -1 or -4 m/s. Lateral actions are to stay on lane, change to left lane, or change to right lane. In this paper, we compare the performance of this baseline architecture with our new architecture described below.

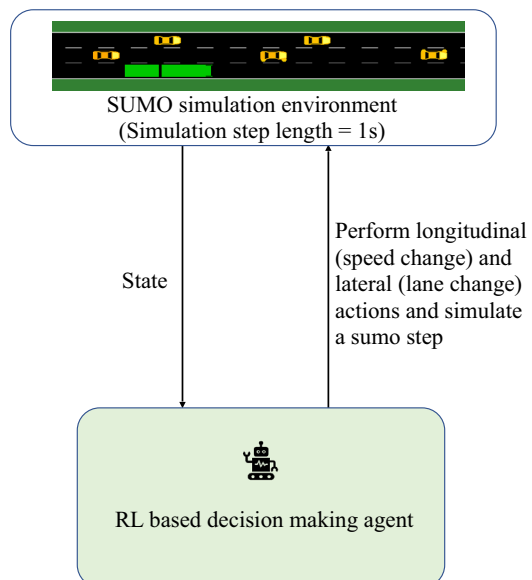
2.3.2 New architecture

Figure 3 shows the new architecture diagram for the automated truck driving framework. It has mainly three components; a high level decision making agent based on RL, a low level controller for longitudinal control and a low level controller for lateral control. The RL agent chooses a longitudinal action or a lateral action based on the current state of the SUMO environment.

The action space is defined below.

1. Set short time gap with leading vehicle (1 s)
2. Set medium time gap with leading vehicle (2 s)
3. Set long time gap with leading vehicle (3 s)
4. Increase desired speed by 1 m/s
5. Decrease desired speed by 1 m/s
6. Maintain current desired speed and time gap

Fig. 2 Overview of the baseline architecture



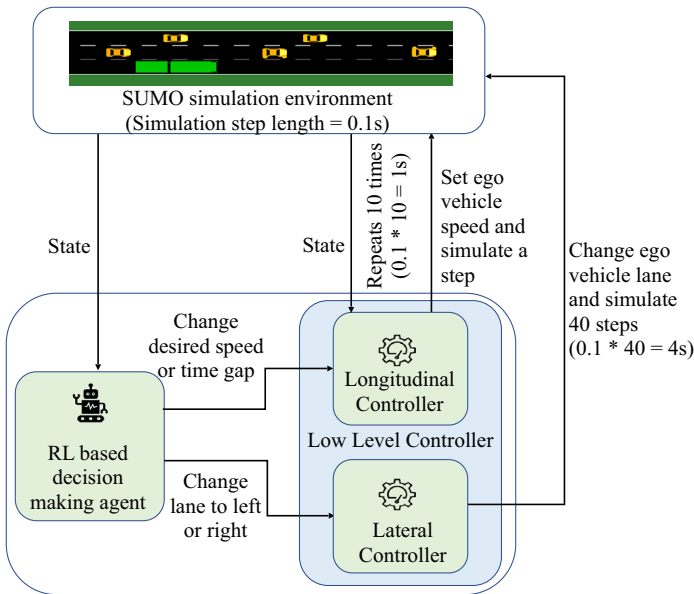


Fig. 3 Overview of the new architecture

- 7. Change lane to left
- 8. Change lane to right

The longitudinal action can be one of the following: set desired time gap (short, medium, long), increment or decrement desired speed or maintain the current time gap and desired speed. When RL agent chooses one of these actions, the longitudinal controller will compute acceleration of ego vehicle based on Intelligent Driver Model (IDM) (Treiber et al. 2000) given by,

$$\dot{v}_\alpha = \frac{dv_\alpha}{dt} = a \left(1 - \left(\frac{v_\alpha}{v_0} \right)^\delta - \left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{s_\alpha} \right)^2 \right), \tag{7}$$

$$s^*(v_\alpha, \Delta v_\alpha) = s_0 + v_\alpha T + \frac{v_\alpha \Delta v_\alpha}{2\sqrt{ab}}$$

where α is the ego vehicle and $\alpha - 1$ is the leading vehicle. v denotes the velocity and l denotes the length of the vehicle. $s_\alpha := x_{\alpha-1} - x_\alpha - l_{\alpha-1}$ is the net distance and $\Delta v_\alpha := v_\alpha - v_{\alpha-1}$ is the velocity difference. v_0 (desired velocity), s_0 (minimum spacing), T (desired time gap), a (maximum acceleration), and b (comfortable braking deceleration) are model parameters.

Here IDM uses the latest desired speed and time gap set by the RL agent. It computes a new acceleration for the truck and sets the resulting speed in SUMO every 0.1s. This process continues for a total duration of 1 s, after which the RL agent chooses the next high level action. Note that in the new architecture, the ACC mode is always activated for the truck by construction.

The lateral action is to change the truck’s course to left or right lane. When the RL agent chooses a lateral action, the lateral controller initiates the lane change. Lane change is performed using the default LC2013 lane change model (Erdmann 2014) in SUMO. The lane width is set to 3.2m and the lateral speed of the truck is set to 0.8m/s. Hence, in total, it takes 4s to complete a lane change action, which corresponds to 40 SUMO time steps of 0.1s duration. Following this, the RL agent chooses the next high level action.

2.4 Total cost of operation

The TCOP of a truck encompasses various expenses incurred during its operation such as energy cost, driver cost and insurance cost. We have designed a TCOP-centric reward function to provide the agent with more realistic goals similar to those of a human truck driver and to motivate the agent to learn safe and cost-effective actions. The reward at each time step is given by (8). Here, we consider realistic cost or revenue values (in euros) for each component to see if the agent can learn a safe and cost effective driving strategy with this reward function.

$$r(t) = -C_{el} e_t - C_{dr} \Delta t - I_l P_l - I_c P_c W_c - I_{nc} P_{nc} W_{nc} - I_o P_o W_o + I_{tar} R_{tar} W_{tar} \tag{8}$$

C_{el} is the electricity cost, e_t is the electricity consumed at time step t , C_{dr} is the driver cost and Δt is the duration of a time step. The electricity consumed during the time step t (e_t) is calculated as,

$$e_t = f_t v_t \Delta t, \tag{9}$$

where f_t , force at time step t is given by,

$$f_t = m a_t + \frac{1}{2} C_d A_f \rho_{air} v^2 + m g C_r + m g \sin(\arctan(\frac{slope}{100})) \tag{10}$$

here m is the mass of the vehicle, C_d is the coefficient of air drag, A_f is the frontal area, ρ_{air} is the air density, C_r is the coefficient of rolling resistance, g is the acceleration due to gravity and a is the acceleration of the vehicle at time step t . The slope of the road is set to be 0 in this study. The parameter values used are given in Table 2.

In the reward function (8), the penalties are defined as the average cost incurred during hazardous situations. The average cost considered here is the *own risk payment* which is the portion of an insurance claim made by the vehicle owner or the insured for any loss and or damage that occurs when submitting a claim. Similarly, the reward R_{tar} denotes the revenue that can be achieved by the truck when it completes the target. Revenue is computed as the total expected cost with 20% profit in an ideal scenario where the truck drives with an average speed of 22m/s and zero acceleration. The total length of travel is 2200 m and hence the time to reach the target will be 100s. Based on (9) and parameter values from Table 2, the total energy consumed can be computed as 1.85 kwh. Then, the total energy cost will

Table 2 Parameter values used in TCOP based reward function

Parameter	Value
P_l	0.1
P_c	1000 euros
P_{nc}	1000 euros
P_o	1000 euros
R_{tar}	2.78 euros
C_{el}	0.5 euro per kwh
C_{dr}	50 euro per hour
Δt	1 s
m	40000 kg
C_d	0.36
A_f	10 m^2
ρ_{air}	1.225 kg/m^3
g	9.81 m/s^2
C_r	0.005

be $1.85 \text{ kwh} \times 0.5 = 0.925$ euros. The total driver cost for 100s will be 1.39 euros and the total cost becomes $0.925 + 1.39 = 2.315$ euros. Further, adding the 20% profit gives the net revenue 2.78 euros, which is used in the reward function along with a weight W_{tar} . We also added weight components W_c , W_{nc} , W_o for the penalties for collision, near collision and driving outside the road condition respectively.

2.5 Curriculum reinforcement learning

Curriculum Reinforcement Learning (CRL) is a training strategy where the model is gradually exposed to increasingly complex tasks or difficult examples over time. This approach is based on the intuition that starting with simpler tasks allows the model to build foundational knowledge and gradually progress to more challenging tasks, which can help improve learning efficiency.

Effectiveness of curriculum learning has been investigated in a number of problems with complex tasks related to autonomous driving, though different from our setting. The study in Anzalone et al. (2022) introduces multi-stage learning in complex driving environments. Different stages include a diverse set of starting locations, varying weather conditions, and dense traffic scenarios. The action space consists of accelerator or brake values and steering angles. The reward function penalizes collisions, following the wrong route, and exceeding speed limits, guiding the agents towards safe and efficient driving behaviors. The papers (Liu et al. 2023) and Song et al. (2021) utilize a curriculum learning approach for overtaking in autonomous driving. Liu et al. (2023) employs a two-stage successive learning progression, where agents first learn to drive as fast as possible and then master the skill of overtaking efficiently. On the other hand, (Song et al. 2021) introduces a three-stage curriculum learning process where agents first learn to race, then focus on overtaking maneuvers, and finally refine their skills to race at high speeds while avoiding collisions. Both of these works demonstrate that curriculum learning boosts convergence and contributes to a better final policy.

The above mentioned studies do not address the problem of strategic decision making for optimizing cost, efficiency and safety at the same time. We address this by combining

curriculum learning with RL in the new system architecture and investigate its effectiveness in training the RL agent with the complex reward function based on TCOP. Here, the curriculum for training the RL agent is designed as follows:

- **Curriculum-1:** Learning longitudinal and lateral control without collision/driving outside the road and by reducing driver cost
- **Curriculum-2:** Learn to minimize energy cost
- **Curriculum-3:** Learn to reach the destination successfully within maximum steps

In this CRL approach, we update the reward function in each curriculum whereas the RL environment, state and action space remain the same. In this work, we have redefined TCOP based reward function as shown in Eq. 11 by removing the penalty for lane change and the weight components such that the reward function only contains actual cost and revenue values during a truck’s operation. Training an agent with realistic reward functions can lead to policies that are more aligned with human expectations.

$$r(t) = -C_{el} e_t - C_{dr} \Delta t - I_c P_c - I_{nc} P_{nc} - I_o P_o + I_{tar} R_{tar} \tag{11}$$

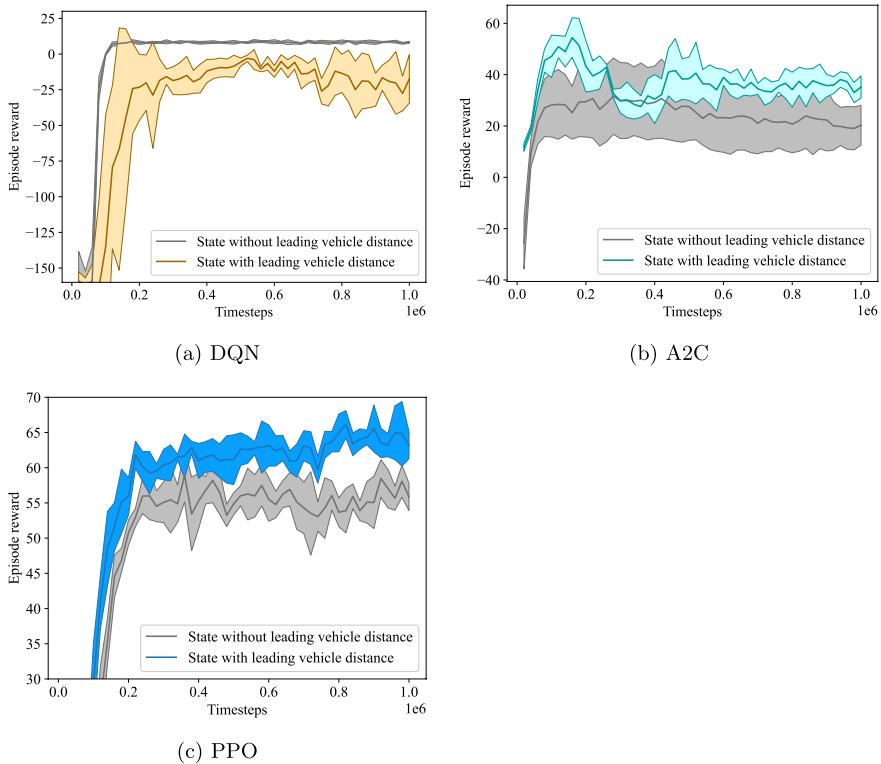


Fig. 4 Comparison of episode rewards in baseline architecture with and without leading vehicle distance in state space

here, the notations and values are the same as in Eq. 8. This reward function is divided into smaller components and added to the reward functions $r_1(t)$, $r_2(t)$, $r_3(t)$ defined in Eq. 12 for curriculum-1, curriculum-2 and curriculum-3 respectively.

$$\begin{aligned} r_1(t) &= -I_c P_c - I_{nc} P_{nc} - I_o P_o - C_{dr} \Delta t \\ r_2(t) &= r_1(t) - C_{el} e_t \\ r_3(t) &= r_2(t) + I_{tar} R_{tar} \end{aligned} \quad (12)$$

here, the negative reward components for drive cost and energy cost contradict to each other as driver cost motivates the agent to drive faster while energy cost motivates the agent to drive slower. This will lead to challenges in learning an optimal policy as shown in the experiments section. Therefore, we tuned the reward function for curriculum-1 and curriculum-2 to normalize these two components with the distance travelled per time step Δd as shown in Eq. 13.

$$\begin{aligned} r_1(t) &= -I_c P_c - I_{nc} P_{nc} - I_o P_o - \frac{(C_{dr} \Delta t)}{\Delta d} \\ r_2(t) &= r_1(t) - \frac{(C_{el} e_t)}{\Delta d} \end{aligned} \quad (13)$$

3 Experiments

This section presents the results from different experiments conducted on the SUMO platform using the baseline and new architectures with different RL algorithms. The experiments are conducted on a linux cluster that consists of 28 CPUs with the model *Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz*. The average time elapsed for training the new architecture with DQN, A2C and PPO for $1e^6$ timesteps are 12528 s (3 h 28min), 13662 s (3 h 47min), 14124 s (3 h 55min) respectively.

3.1 Performance improvement based on selection of states

First, we show how the performance can be improved by adding relevant features to the observation space. As mentioned in section II, the state space includes position, speed, and lane change information of the ego vehicle and the surrounding vehicles. In this experiment,

Table 3 Evaluation of baseline architecture using PPO with and without leading vehicle distance in state space

Evaluation metric	Without leading distance	With leading distance
Reached target successfully	61%	70.6%
Driven successfully, but not reached target within maximum steps	0%	0%
Terminated by collision or driving outside road	39%	29.4%
Average speed	18.17 m/s	19.43 m/s
Average distance travelled	1503.93 m	1667.86 m
Average executed steps	88.43	89.65

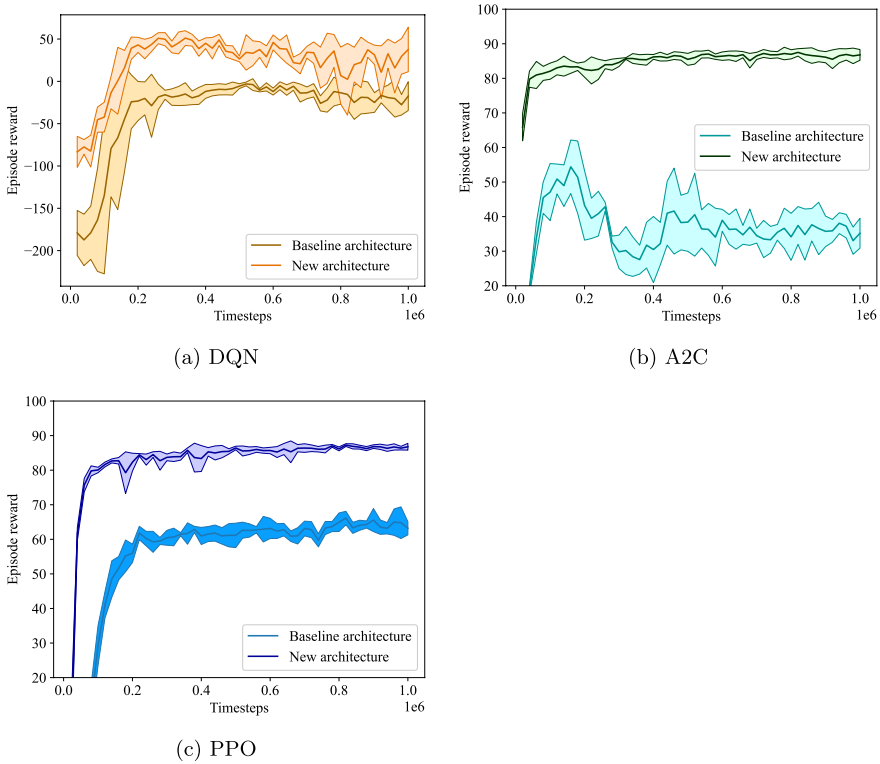


Fig. 5 Comparison of average episodic reward in the baseline and new architectures with different RL agents

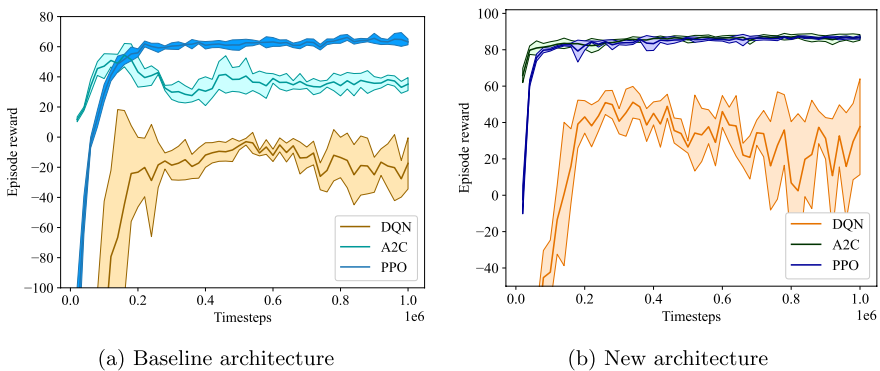
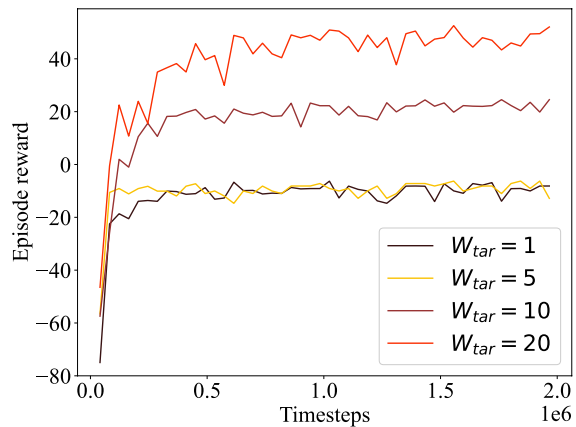


Fig. 6 Performance comparison of different RL agents

we compare the performance of baseline architecture by adding the distance to the leading vehicle as an explicit feature in the above mentioned state space. The basic reward function is used in both cases. The learning curves of average reward (over 5 realizations) using different RL algorithms are shown in Fig. 4. The results from PPO and A2C show that the

Table 4 Evaluation of baseline and new architectures using PPO

Evaluation metric	Baseline architecture	New architecture
Reached target successfully	70.6%	97.8%
Driven successfully, but not reached the target within maximum steps	0%	0.6%
Terminated by collision or driving outside the road	29.4%	1.6%
Average speed	19.43 m/s	18.56 m/s
Average distance travelled	1667.86 m	2178.37 m
Average executed steps	89.65	128.75

Fig. 7 Learning of TCOP-based rewards in the new architecture with PPO using different W_{tar} values and $W_c, W_{nc}, W_o = 0.1$ 

average reward has improved by adding distance to leading vehicle as a feature in the state. Results from DQN show the opposite trend where the average reward for state without leading vehicle distance is higher. However, the average reward values are very low for DQN in both state spaces, and the experiments in upcoming sections also show that DQN is not able to achieve good performance for this specific problem.

Table 3 shows the validation results for PPO algorithm with different evaluation metrics (average of 5 validations with 100 episodes each). The collision rate is considerably reduced and the average speed of the ego vehicle is slightly increased. Note that this distance was already available in the observation space among the properties of surrounding vehicles. Our results show that explicitly adding this as a separate feature helps the agent to learn better how to control the speed w.r.t. the leading vehicle and avoid forward collisions.

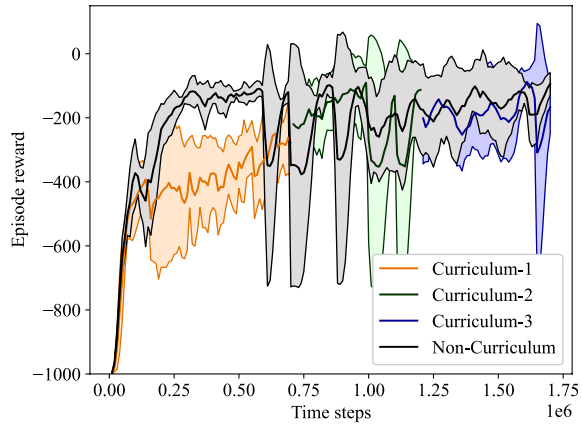
3.2 Performance comparison of the two architectures

In this experiment, we compare the performance of our new architecture shown in Fig. 3 with the baseline framework shown in Fig. 2 using different RL algorithms. For both architectures, we use the observation space including leading vehicle distance and the basic reward function. However, the new architecture introduces a low level controller based on a physical model to perform speed control actions. Figure 5 shows the comparison of average episode rewards in both architectures for different RL algorithms. It can be seen that the new architecture outperforms the baseline regardless of the chosen RL algorithm. Further, Fig. 6

Table 5 Evaluation of new architecture with TCOP based reward function, using different W_{tar} values and $W_c, W_{nc}, W_o = 0.1$

Evaluation metric	$W_{tar} = 1$	$W_{tar} = 5$	$W_{tar} = 10$	$W_{tar} = 20$
Reached target successfully	2.2%	13.6%	68.2%	99.2%
Driven successfully, but not reached the target within maximum steps	95.2%	83.6%	29.6%	0%
Terminated by collision or driving outside road	2.6%	2.8%	2.2%	0.8%
Average speed	1.75 m/s	2.78 m/s	11.16 m/s	18.36 m/s
Average distance travelled	615.9 m	848.2 m	1864.7 m	2197.4 m
Average executed steps	482	455	265	122
Average energy cost	0.17 euros	0.28 euros	1.1 euros	2.05 euros
Average driver cost	6.7 euros	6.32 euros	3.68 euros	1.69 euros
Average tcop	6.87 euros	6.6 euros	4.78 euros	3.74 euros

Fig. 8 Learning of TCOP based reward function without normalization using CRL and non-CRL approaches



compares performance of different RL agents in each architecture. For both cases, DQN obtains lowest average rewards. PPO obtains highest average reward in baseline architecture whereas both PPO and A2C shows similar performance in new architecture.

As PPO gives consistent performance in both architectures, it is used for further validation. The validation results in Table 4 clearly show that the new architecture has improved the performance compared to the baseline. The collision rate is reduced immensely from 29.4% to 1.6%. The slight decrease in the average speed of the new framework is negligible when considering the improvement in safety. The reduced collision rate also explains why there is an increment in the average travelled distance and average executed steps.

Table 6 Evaluation of agent trained with TCOP based reward function without normalization using CRL and non-CRL approaches

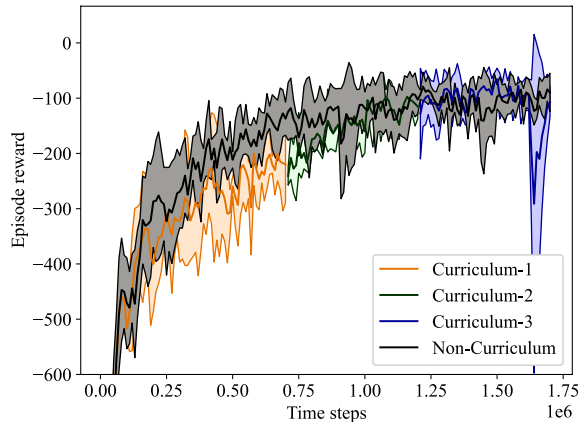
Evaluation metric	Curriculum-1	Curriculum-2	Curriculum-3	Non-curriculum
Reached target successfully	44.80%	10.35%	7.15%	2.50%
Terminated by collision or driving outside road	13.45%	8.80%	9.60%	6.20%
Driven successfully, but not reached the target within maximum steps	41.75%	80.85%	83.25%	91.30%
Average distance travelled	1269.80 m	837.52 m	739.12 m	603.66 m
Average executed steps	265.91	431.43	438.64	463.48
Average speed	11.38 m/s	3.80 m/s	2.99 m/s	2.14 m/s
Average energy cost	1.23 euros	0.38 euros	0.27 euros	0.20 euros
Average driver cost	3.7 euros	6 euros	6.1 euros	6.44 euros
Average t_{cop}	4.93 euros	6.38 euros	6.37 euros	6.64 euros

3.3 Performance with TCOP based reward

3.3.1 Comparison with different weight values

In this analysis, we apply the TCOP-based reward function in Eq. 8 to the new architecture in order to assess whether the agent can successfully acquire a driving strategy that is both safe and cost-effective. The episodic reward is evaluated with varying weights W_{tar} assigned to the target completion R_{tar} , as depicted in Fig. 7. As mentioned in Section II, in an ideal scenario, the cost would amount to 2.315 euros, while the revenue (R_{tar}) would be 2.78 euros. Consequently, in this ideal case, the episodic reward would equal $2.78 - 2.315 = 0.46$ when W_{tar} is 1. This value is relatively small, making it challenging

Fig. 9 Learning of TCOP based reward function with normalization using CRL and non-CRL approaches



for the agent to learn how to reach the target. Furthermore, the penalties for hazardous situations such as collision, near collision and driving outside the road are 1000 euros which is comparatively very high. This leads to a risk for the agent to focus only on avoiding such situations and do not care about reaching the target, for example by driving with very low speed. This issue can be overcome by carefully choosing the weight values for the penalty and revenue components. We observed good results by keeping the weight value for penalties $W_c, W_{nc}, W_o = 0.1$ and increasing the weight value for revenue W_{tar} to 20 as illustrated in Fig. 7. From Table 5, it becomes evident that increasing the weight W_{tar} assists the agent in learning to drive at higher speeds and successfully reach the target. Notably, as the weight W_{tar} is increased, the total cost of operation—comprising energy cost and driver cost—is minimized to 3.74 euros. This cost is lower than the cost incurred when the ego vehicle was entirely controlled by SUMO, using the default Krauss car following model and LC2013 lane change model in the same traffic environment, which resulted in a cost of 3.85 euros.

3.3.2 Comparison with and without reward normalization in CRL and non-CRL

In these experiments, instead of adding weights to the reward components, we investigate whether the agent can learn an optimal policy with the modified TCOP reward function as in Eq. 11 that contains only the actual costs and revenue values. We compare the performance with and without CRL approach using PPO algorithm. PPO was chosen because it proved to give better performance in earlier experiments. To improve exploration, we used an adaptive entropy coefficient of 0.01 which gradually decreases to 0.001 whereas default values in stable-baselines3 are used for all other hyperparameters.

Figure 8 shows the episodic reward training curve for the RL agent trained with curriculum learning and compares it to the training curve without using curriculum learning. Here, the agent is trained with curriculum learning using the reward functions without normalization given in Eq. 12 where curriculum-1 is trained for $7e^5$ time steps and curriculum-2 and curriculum-3 are trained for $5e^5$ time steps each. Non-curriculum learning results are obtained from RL agent trained with the reward function Eq. 11 for $17e^5$ time steps. There are noticeable drops in the reward curve at certain points during training in both curriculum and non-curriculum learning. The validation results in Table 6 also show a poor success rate.

Table 7 Evaluation of agent trained with TCOP based reward function with normalization using CRL and non-CRL approaches

Evaluation metric	Curriculum-1	Curriculum-2	Curriculum-3	Non-curriculum
Reached target successfully	57.45%	49.15%	73.50%	73.25%
Terminated by collision or driving outside road	8.25%	4.45%	2.90%	3.50%
Driven successfully, but not reached the target within maximum steps	34.30%	46.40%	23.60%	23.25%
Average distance travelled	1569.99 m	1382.20 m	1794.74 m	1770.03 m
Average executed steps	252.24	294.00	210.22	206.94
Average speed	12.27 m/s	10.38 m/s	14.18 m/s	14.42 m/s
Average energy cost	1.33 euros	1.08 euros	1.68 euros	1.62 euros
Average driver cost	3.50 euros	4.08 euros	2.92 euros	2.88 euros
Average tcop	4.83 euros	5.17 euros	4.60 euros	4.50 euros

Here, the evaluation is performed by averaging the validation results from 5 trained models for curriculum and non-curriculum approaches.

It can be seen that the average speed of the ego vehicle drops down largely after training with curriculum-2 when the penalty for energy cost is introduced, which also leads to a reduction in the success rate. This motivates us to tune the reward function by normalizing driver cost and energy cost by distance travelled per time step. Figure 9 and Table 7 show the results of the agent trained with this normalized reward function in Eq. 13 and corresponding non-curriculum learning. Here, the reduction in speed from curriculum-1 to curriculum-2 is smaller and this reduction is recovered when the reward for target completion is added to the next curriculum. Consequently, the success rate is higher in last curriculum compared to the previous case. Figure 10 compares the training curves with and without

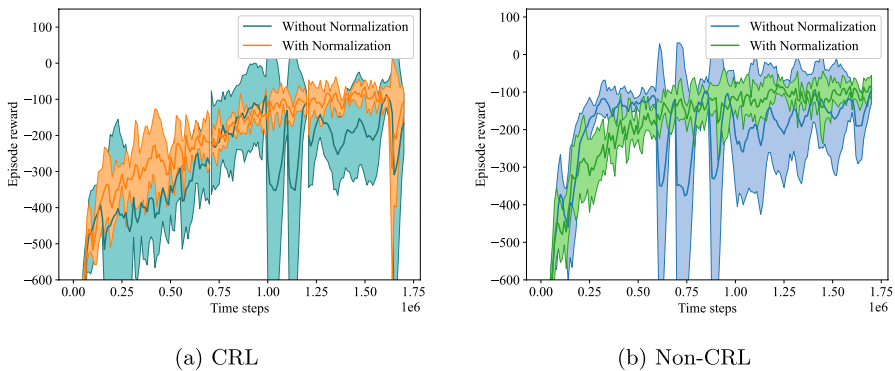


Fig. 10 Performance comparison with and without normalization in TCOP based rewards

reward normalization where we could observe that the average reward is comparatively higher with normalization in both CRL and non-CRL approaches. The training curve with normalization also shows a more stable and steady improvement in rewards, even though the drop in the last curriculum persists. Overall, the normalized reward function significantly improves the results for both CRL and non-CRL approaches as shown explicitly in Table 8. We could observe significant improvement in success rate, average speed and TCOP. However, the results obtained for curriculum and non-curriculum approaches are very similar which demonstrates that there is no particular advantage in using curriculum learning in this RL setting.

4 Conclusion

We implemented an RL framework for tactical decision making of autonomous trucks in a highway environment by integrating RL with low-level controllers. Our results demonstrate that training the agent to focus on high level decisions, such as maintaining a time gap with the leading vehicle, while leaving the low level speed control to a physics-based controller, accelerates learning and improves overall performance. Additionally, we explored the effectiveness of training the agent with realistic rewards and penalties using a multi-objective TCOP based reward function. We study this setting with different approaches, by adding weights to reward components, by normalizing the reward components and by using curriculum learning techniques. We could observe that reshaping the reward function with weights or normalization significantly improves the performance whereas CRL shows comparable results with non-CRL approach.

An interesting future direction would be to explore transfer learning methods for generalizing tactical decision-making to diverse traffic scenarios such as uphill, downhill, merging traffic etc. In this context, we believe that the pre-trained models from simple highway environments can serve as a strong foundation while adapting to new environments with different dynamics and constraints.

Table 8 Comparison of validation results with and without normalization in TCOP based rewards in CRL and non-CRL approaches

Evaluation metric	CRL		Non-CRL	
	Without normalization	With normalization	Without normalization	With normalization
Reached target successfully	7.15%	73.50%	2.50%	73.25%
Terminated by collision or driving outside road	9.60%	2.90%	6.20%	3.50%
Driven successfully, but not reached the target within maximum steps	83.25%	23.60%	91.30%	23.25%
Average distance travelled	739.12 m	1794.74 m	603.66 m	1770.03 m
Average executed steps	438.64	210.22	463.48	206.94
Average speed	2.99 m/s	14.18 m/s	2.14 m/s	14.42 m/s
Average energy cost per meter	0.00036 euros	0.00094 euros	0.00033 euros	0.00092 euros
Average driver cost per meter	0.0082 euros	0.0016 euros	0.0107 euros	0.0016 euros
Average tcop per meter	0.0086 euros	0.0026 euros	0.011 euros	0.0025 euros

Acknowledgements This work was partially supported by Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation. We would like to thank Nikolce Murgovski, Erik Börve and Stefan Börjesson for valuable discussions.

Author contributions All authors have contributed significantly to the conceptualization, design, writing and editing the manuscript. Deepthi Pathare has implemented the RL framework, conducted the experimental studies, and written most parts of the paper, under the supervision of Leo Laine and Morteza Haghir Chehreghani.

Funding Open access funding provided by Chalmers University of Technology.

Data availability No datasets were generated or analysed during the current study.

Code availability The source code is available at: <https://github.com/deepthi-pathare/Autonomous-truck-su-mo-gym-env>.

Declarations

Conflict of interest There is no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anzalone L, Barra P, Barra S, Castiglione A, Nappi M (2022) An end-to-end curriculum learning approach for autonomous driving scenarios. *IEEE Trans Intell Transp Syst* 23(10):19817–19826. <https://doi.org/10.1109/TITS.2022.3160673>
- Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning, pp 41–48
- Das LC, Won M (2021) Saint-acc: safety-aware intelligent adaptive cruise control for autonomous vehicles using deep reinforcement learning. In: ICML
- De Jong G, Gunn H, Walker W (2004) National and international freight transport models: an overview and ideas for future development. *Transp Rev* 24(1):103–124
- Desjardins C, Chaib-draa B (2011) Cooperative adaptive cruise control: a reinforcement learning approach. *IEEE Trans Intell Transp Syst* 12(4):1248–1260
- Erdmann J (2014) Lane-changing model in sumo. In: Proceedings of the SUMO2014 modeling mobility with open data. Reports of the DLR-institute of transportation systems proceedings,
- Grislis A (2010) Longer combination vehicles and road safety. *Transport* 25(3):336–343
- Haarnoja T, Zhou A, Abbeel P, Levine S (2018) Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor
- Hoel CJ, Wolff K, Laine L (2020) Tactical decision-making in autonomous driving by reinforcement learning with uncertainty estimation. In: Intelligent vehicles symposium. IEEE

- Jiménez F, Naranjo JE, Anaya JJ, García F, Ponz A, Armingol JM (2016) Advanced driver assistance system for road environments to improve safety and efficiency. *Transp Res Procedia* 14
- Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, Pérez P (2021) Deep reinforcement learning for autonomous driving: a survey. *Trans Intell Transp Syst* 23(6):4909–4926
- Konda V, Tsitsiklis J (1999) Actor-critic algorithms. In: Solla, S., Leen, T., Müller, K. (eds.) *Advances in neural information processing systems*
- Krauss S, Wagner P, Gawron C (1997) Metastable states in a microscopic model of traffic flow. *Phys Rev E* 55(5):5597
- Lin Y, McPhee J, Azad NL (2021) Comparison of deep reinforcement learning and model predictive control for adaptive cruise control. *IEEE Trans Intell Veh* 6(2):221–231
- Liu J, Li H, Yang Z, Dang S, Huang Z (2023) Deep dense network-based curriculum reinforcement learning for high-speed overtaking. *IEEE Intell Transp Syst Mag* 15(1):453–466. <https://doi.org/10.1109/IMITS.2022.3174410>
- Lopez PA, Behrisch M, Bieker-Walz L, Erdmann J, Flötteröd YP, Hilbrich R, Lücken L, Rummel J, Wagner P, Wießner E (2018) Microscopic traffic simulation using sumo. In: *International conference on intelligent transportation systems*
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller MA, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. *Nature* 518:529–533
- Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: Balcan MF, Weinberger KQ (eds.) *Proceedings of the 33rd international conference on machine learning proceedings of machine learning research*, vol. 48, pp 1928–1937. PMLR, New York, New York, USA. <https://proceedings.mlr.press/v48/mniha16.html>
- Moreno G, Nicolazzi LC, Vieira RDS, Martins D (2018) Stability of long combination vehicles. *Int J Heavy Veh Syst* 25
- Narvekar S, Peng B, Leonetti M, Sinapov J, Taylor ME, Stone P (2020) Curriculum learning for reinforcement learning domains: a framework and survey. *J Mach Learn Res* 21(1):7382–7431
- Nilsson P, Laine L, Sandin J, Jacobson B, Eriksson O (2018) On actions of long combination vehicle drivers prior to lane changes in dense highway traffic - a driving simulator study. *Transp Res F: Traffic Psychol Behav* 55:25–37
- Pathare D, Laine L, Chehreghani MH (2023) Improved tactical decision making and control architecture for autonomous truck in sumo using reinforcement learning. In: *2023 IEEE international conference on big data (BigData)*, pp 5321–5329. <https://doi.org/10.1109/BigData59044.2023.10386803>
- Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N (2021) Stable-baselines3: reliable reinforcement learning implementations. *J Mach Learn Res* 22(268):1–8
- Sallab AE, Abdou M, Perot E, Yogamani S (2017) Deep reinforcement learning framework for autonomous driving. *arXiv preprint arXiv:1704.02532*
- Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms
- Shalev-Shwartz S, Shammah S, Shashua A (2016) Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*
- Shaout A, Colella D, Awad S (2011) Advanced driver assistance systems-past, present and future. In: *International computer engineering conference. IEEE*
- Song Y, Lin H, Kaufmann E, Durr P, Scaramuzza D (2021) Autonomous overtaking in gran turismo sport using curriculum reinforcement learning, pp 9403–9409. <https://doi.org/10.1109/ICRA48506.2021.9561049>
- Sutton RS, Barto AG (2018) *Reinforcement learning: an introduction*, 2nd edn
- Svensson HG, Tyrchan C, Engkvist O, Chehreghani MH (2023) Utilizing reinforcement learning for de novo Drug Design 13(7):4811–4843
- Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 62(2):1805
- Xiao L, Gao F (2010) A comprehensive review of the development of adaptive cruise control systems. *Veh Syst Dyn* 48(10):1167–1192
- Yang D, Qiu X, Yu D, Sun R, Pu Y (2015) A cellular automata model for car-truck heterogeneous traffic flow considering the car-truck following combination effect. *Phys A: Stat Mech Appl* 424
- Yang J, Liu X, Liu S, Chu D, Lu L, Wu C (2020) Longitudinal tracking control of vehicle platooning using ddpq-based pid. In: *2020 4th CAA international conference on vehicular control and intelligence (CVCI)*
- Zhao D, Wang B, Liu D (2013) A supervised actor-critic approach for adaptive cruise control. *Soft Comput* 17(11):2089–2099