Thesis for The Degree of Licentiate of Engineering

Efficient Tactical Decision Making for Trucks in Highway Traffic with Deep Reinforcement Learning

Deepthi Pathare

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2026

Efficient Tactical Decision Making for Trucks in Highway Traffic with Deep Reinforcement Learning

Deepthi Pathare

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Division of Data Science and AI
Machine Learning & Decision Making Lab
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

To my family

Efficient Tactical Decision Making for Trucks in Highway Traffic with Deep Reinforcement Learning

Deepthi Pathare

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

This thesis investigates tactical decision making for autonomous heavy-duty trucks in highway traffic using deep reinforcement learning, with a particular emphasis on optimizing safety, efficiency and costs. The key aspects of decision making include Adaptive Cruise Control (ACC) and lane changes, which strongly influence energy consumption, travel time, and traffic interactions. To support a systematic study of this problem, we develop a scalable traffic model on a simulation platform, providing a controlled and extensible environment for autonomous truck driving in multi-lane highways.

We propose a hierarchical control architecture in which reinforcement learning is used for high-level tactical decision making, while low-level tactical actions are handled by physics-based controllers. This separation is found to improve the performance by reducing safety risks and facilitates the integration of learning-based decision making with established control methods. A realistic reward function is designed to jointly capture safety, efficiency, and operational costs, and advanced training strategies such as curriculum learning are investigated to handle conflicting objectives within a scalarized framework.

We further explore a multi-objective reinforcement learning formulation to explicitly represent trade-offs between competing objectives, enabling the learning of interpretable Pareto frontiers. The results demonstrate that learning-based tactical decision making policies can achieve meaningful trade-offs between safety and various operational costs in abstracted highway scenarios, and that multi-objective formulations provide valuable insight into the structure of these trade-offs. Overall, this work contributes to methodological foundations and evaluation tools for economically meaningful and extensible learning-based tactical decision making for heavy-duty trucks.

Keywords

Autonomous driving, Tactical Decision making, Deep Reinforcement Learning, Multi-Objective Reinforcement Learning

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **Deepthi Pathare**, Leo Laine, Morteza Haghir Chehreghani., *Tactical decision making for autonomous trucks by deep reinforcement learning with total cost of operation based reward. Artificial Intelligence Review, Volume 59, 2025*

[**Paper II**] **Deepthi Pathare**, Leo Laine, Morteza Haghir Chehreghani., *Multi-Objective Reinforcement Learning for Efficient Tactical Decision Making for Trucks in Highway Traffic. arXiv preprint, arXiv:2601.18783. Under review, 2026*

Other publications

The following publications were published during my PhD studies. However, they are not appended to this thesis, due to contents overlapping that of appended publications.

[**a**] **Deepthi Pathare**, Leo Laine, Morteza Haghir Chehreghani, *Improved tactical decision making and control architecture for autonomous truck in sumo using reinforcement learning.* IEEE International Conference on Big Data (BigData), 5321-5329, 2023

(Paper I is an extended version of this paper)

Acknowledgment

I would like to express my sincere gratitude to my supervisor, Morteza Haghir Chehreghani, for his continuous guidance and encouragement throughout this journey. His support has been invaluable to the progress of this work. I am equally grateful to my industrial supervisor, Leo Laine, for his insights and steady support, which have played an important role in shaping my research. I sincerely thank my co-supervisor, Nikolce Murgovski, for valuable discussions and feedback during the course of my work. I also thank my examiner, Patrik Jansson, for his support during this journey.

I gratefully acknowledge Volvo Group and Chalmers University of Technology for providing me an opportunity to pursue this PhD. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation and I gratefully acknowledge them for enabling this research.

I would like to thank Erik, Stefan, and Markus at Volvo for their collaboration and valuable research discussions. I would also like to thank my manager, Daniel, and my colleagues in the Sensor Fusion and Maps team for creating a supportive work environment. Furthermore, I would like to thank my previous managers, Rached and Sofia, and all colleagues within the Division of Safe Efficient Driving and Services for their support throughout this journey.

I am grateful to Yossra for being such a good friend. I am also very thankful to Emma, Jack, Kilian, Valter, Hampus, Mengyuan, and all other fellow PhD students at Chalmers for creating an enjoyable work environment. I acknowledge the faculty members, administrative staff, and other supporting personnel at the Division of Data Science and AI for their support.

I would like to thank my family and friends for their immense support during this journey. I am deeply grateful to Amma and Achan for always believing in me. I am also thankful to my sisters, Divya, Deepa, and Preetha, my brothers-in-law, Jayan and Renish, and my best friends, Nayana and Jimi, for all the joy and encouragement. Finally, Hari, thank you for being there, always, constantly motivating me and cheering me up, and for celebrating even my smallest achievements with so much joy.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

Road freight transport plays a critical role in modern societies, enabling economic activities by facilitating the movement of goods over long distances. Heavy-duty vehicles (HDVs), such as long-haul trucks, account for a substantial share of freight transport and are indispensable to supply chains worldwide. In Europe, road freight transport carries more than 70% of inland freight, with HDVs contributing significantly to economic productivity but also to energy consumption, greenhouse gas emissions, and traffic congestion [1]. Despite representing a relatively small fraction of the total vehicle fleet, heavy-duty trucks are responsible for a large share of fuel consumption and CO_2 emissions due to their large mass and high aerodynamic drag [2], [3]. Moreover, even minor control errors in HDV operation can lead to severe crashes, the majority of which are caused by human factors such as fatigue, inattention, or suboptimal decision making [4], [5]. Improving the operational efficiency and safety of HDVs is therefore of high societal, economic, and environmental importance.

Automation of driving tasks offers a promising avenue to mitigate these risks by supporting drivers in complex traffic scenarios, reducing human error, and improving both safety and efficiency [6]. Traditional autonomous driving systems have primarily relied on rule-based and optimization-based architectures. Rule-based approaches encode expert knowledge through hand-crafted logic and safety constraints, offering interpretability, predictable behavior, and real-time performance. However, their reliance on predefined rules limits adaptability and often results in overly conservative behavior in complex or rapidly changing traffic [7], [8]. Optimization-based methods, particularly Model Predictive Control (MPC), provide a principled framework for planning safe and feasible maneuvers by optimizing cost functions subject to vehicle dynamics and constraints [9], [10]. However, they depend on accurate models and can be difficult to scale to traffic scenarios with complex interactions and long planning horizons [11].

In recent years, advances in machine learning (ML), especially Reinforcement Learning (RL) have broadened research on data-driven decision making for autonomous driving in complex and dynamic environments. These approaches aim to learn decision policies that directly map sensory input and state estimates

to maneuver decisions, and have been explored for tasks such as navigation, lane changing, and collision avoidance [12], [13]. However, despite these promising developments, major challenges remain. Notably, when applied to safety-critical domains, most machine learning–based methods lack formal safety guarantees and may exhibit limited interpretability, hindering their validation and acceptance for large-scale deployment in real-world traffic [14], [15].

Most existing studies focus on the automation of passenger cars, and their findings cannot be directly applied to heavy-duty trucks. HDVs introduce additional challenges that significantly increase the complexity of decision making. From a physical perspective, the large mass and inertia of HDVs lead to longer braking distances and slower acceleration, while fuel consumption and efficiency are highly sensitive to speed trajectories, traffic interactions, and road gradients [16], [17]. In addition, HDVs operate under regulatory constraints such as driving and rest time regulations, and operational decisions are tightly coupled to logistics objectives, including delivery deadlines and cost minimization which are rarely considered in passenger car studies. Current autonomous driving solutions for HDVs remain largely limited to constrained and well-defined environments where conservative, rule-based decision strategies dominate [18], [19]. These limitations highlight the need for decision making approaches specifically tailored to HDVs that can handle uncertainty while jointly considering safety, efficiency, and operational objectives.

To address this research gap, this thesis investigates how learning-based decision making frameworks for HDVs can be modeled, evaluated, and optimized under realistic traffic and operational conditions. For this, we develop a scalable simulation environment, tailored to truck driving on highway scenarios. In our setup, we simulate a multi-lane highway segment with heterogeneous vehicles, and the environment allows configurable parameters for vehicle characteristics, traffic conditions, and other scenario settings. While the current implementation focuses on a simple highway scenario, the platform is modular and extensible, allowing the integration of additional vehicle models, cost components, or traffic scenarios in future work. In general, the platform serves as a reproducible and flexible framework to study tactical decision making in heavy-duty trucks.

For commercial truck applications, decision making cannot be evaluated solely based on safety or efficiency metrics. Instead, it must account for the cost of operations, which broadly includes fuel or energy costs, time-related costs such as driver wages and delivery delays, and long-term costs associated with wear and maintenance. This research is motivated by the need for decision making frameworks that are not only safe and efficient but also economically meaningful and adaptable to varying operational requirements. To this end, we design an objective function that captures the economic reality of truck operation and enables a principled assessment of trade-offs between efficiency and performance. Optimizing these objectives at the tactical level is particularly important, as decisions such as speed adaptation and lane selection directly affect energy use, trip duration, and interactions with other vehicles. The different components of operational performance including safety, efficiency, and cost are inherently conflicting, which further motivates exploring this problem as a multi-objective decision making problem. We employ approaches

that can learn a set of optimal driving strategies that represent different trade-offs between different objectives. Such approaches allow for flexible selection of driving behaviors and strategies without retraining, resulting in a robust and adaptive decision making framework for autonomous heavy-duty trucks.

The following research questions are considered in this thesis.

- RQ1: How can tactical decision making for autonomous trucks be formulated within a reinforcement learning framework using a realistic reward function that jointly captures safety, efficiency, and operational costs?
- RQ2: How can physics-based low-level controllers be integrated with a reinforcement learning framework for tactical decision making in autonomous trucks to promote safe and efficient driving behavior?
- RQ3: How can multi-objective reinforcement learning be used to explicitly represent and learn the trade-offs between multiple conflicting objectives in tactical decision making for autonomous trucks?

The thesis is structured as follows. Chapter 2 provides an overview of the relevant background required to understand the appended papers. Chapter 3 summarizes the problem, methodology, results, and contributions of each paper. Chapter 4 concludes the thesis and outlines possible directions for future work. The final part of the thesis contains the two appended papers.

Chapter 2

Background

In this chapter, we provide a background of the topics and concepts discussed in the appended papers.

2.1 Sequential Decision Making Problems

Sequential decision making is a fundamental framework for modeling problems where an agent must make a series of decisions over time, often under uncertainty, to achieve a particular objective. Unlike single-step decision problems, sequential decision making scenarios require consideration of the long-term consequences of actions, as each decision influences not only the immediate outcome but also the future state of the system. Such problems arise in diverse domains, including autonomous vehicles, robotics, healthcare and finance, where decisions must be optimized over time to balance one or more objectives.

A defining feature of many real-world decision making problems is uncertainty. The agent that performs the decision making may have incomplete or noisy information about the system state, limited knowledge of the underlying dynamics, or restricted ability to predict how the environment will respond to actions. Moreover, the effects of decisions are often communicated through feedback signals that summarize performance, rather than through explicit instructions or labeled examples indicating the correct action. As a result, deriving optimal decision rules analytically is often infeasible, and decision making strategies must instead be learned through interaction with the environment.

A sequential decision making problem is commonly described as a Markov Decision Process (MDP). An MDP provides a mathematical framework for modeling decision making problems in stochastic environments and consists of a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where,

- $\mathcal{S}$ is the **state space**, describing all possible configurations of the environment.
- $\mathcal{A}$ is the **action space**, containing all actions available to the agent.

- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the **transition dynamics**, capturing the probability function of state transitions in the environment.
- $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the **reward function**, providing the immediate scalar reward $r_{t+1} = R(s_t, a_t)$ after taking action $a_t \in \mathcal{A}$ in state $s_t \in \mathcal{S}$.
- $\gamma \in [0, 1]$ is the **discount factor**, which encodes the relative importance of immediate versus long term outcomes.

The objective of the agent is to learn a policy $\pi : S \rightarrow A$ which is a mapping of states to actions in a way that maximizes the discounted cumulative return given by,

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.1}$$

where t is the current timestep.

A widely used learning-based approach for estimating effective decision making strategies in such settings is Reinforcement Learning (RL), which is described later in this chapter.

2.2 Tactical Decision Making in Trucks

Decision making in autonomous driving is commonly structured into three layers: strategic, tactical, and operational. The strategic level is responsible for long-term objectives such as route planning and navigation, while the operational level focuses on translating driving decisions into low-level actuator commands. Situated between these layers, the tactical decision making level adapts the strategic plan to the current traffic situation by selecting appropriate maneuvers, such as speed adjustments and lane changes. These decisions play a central role in balancing key objectives including safety, efficiency, and traffic flow.

In this thesis, the tactical decision making framework is further structured in a hierarchical manner to address Adaptive Cruise Control (ACC), which regulates vehicle speed and maintains a safe distance to the leading vehicle, as well as lane changes. The learning agent is designed to select high-level tactical decisions, such as adjusting the desired speed, selecting a desired time gap, or initiating a lane change. The realization of these decisions is handled by physics-based longitudinal and lateral controllers, which perform speed regulation and the execution of lane change maneuvers.

The decision making problem is formulated as an MDP, as described below. Individual papers appended to this thesis and the experiments conducted in those studies may contain slight variations, which are described in the respective papers.

- **State Space ($\mathcal{S}$):** The state is designed to provide the agent with sufficient information about the environment. It includes observations from the ego vehicle and surrounding vehicles within the sensor range, capturing kinematic, lane-level, and signaling information.

Observations of ego vehicle:

1. Longitudinal position
2. Longitudinal speed
3. Lane change state
4. Lane number
5. State of left indicator
6. State of right indicator
7. Distance to the leading vehicle
8. Length of the vehicle
9. Width of the vehicle

Observations of each surrounding vehicle:

1. Relative longitudinal distance to the ego vehicle
2. Relative lateral distance to the ego vehicle
3. Relative longitudinal speed with respect to the ego vehicle
4. Lane change state
5. Lane number
6. State of left indicator
7. State of right indicator
8. Length of the vehicle
9. Width of the vehicle

- **Action Space ($\mathcal{A}$):** The action space consists of high-level tactical decisions that the RL agent can select.

  The discrete action space includes the following options:

  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 the desired speed by 1 m/s
  5. Decrease the desired speed by 1 m/s
  6. Maintain current desired speed and time gap
  7. Change lane to left
  8. Change lane to right

  Currently, the action space is simplified such that the agent selects only one of the discrete maneuvers described above at a time. In future work, we plan to extend the framework to allow simultaneous longitudinal and lateral actions, which would better reflect real-world driving behavior.

- **Transition Dynamics(P):** The transition dynamics is defined by the simulation platform, and is not known to the decision making agent.

- **Reward Function(R):** The design of the reward function plays a central role in guiding the agent to achieve safe and efficient driving behavior. In the context of this thesis, three types of reward formulations are considered, reflecting the progression of our work across the appended papers.

  1. **Basic reward function:** The initial studies evaluated models using a simple, safety-oriented reward function. It encourages the agent to drive at high speed while avoiding hazardous situations and reaching the target successfully. The reward at each time step is defined as,

  $$r_t = \frac{v_t}{max\_v} - I_l P_l - I_c P_c - I_{nc} P_{nc} - I_o P_o + I_{tar} \frac{R_{tar}}{T} \tag{2.2}$$

  Here, v_t is the speed of ego vehicle at timestep t, max_v is the maximum allowed speed of the ego vehicle, I is an indicator function for various conditions (l - lane change, c - collision, nc - near collision, o - driving off-road, tar - reaching the target), and P and R are the associated penalties and rewards. T denotes the total time to reach the target.

  2. **TCOP-based reward function:** The Total Cost of Operation (TCOP) of a truck encompasses various expenses incurred during its operation. To reflect these realistic operational costs in the learning objective, we design a reward function based on TCOP, given by,

  $$\begin{aligned} r_t = -C_{el}\, e_t - C_{dr}\, \Delta t - I_c P_c - I_{nc} P_{nc} \\ - I_o P_o + I_{tar} R_{tar} \end{aligned} \tag{2.3}$$

  Here, we consider realistic cost or revenue values for each component. C_{el} and C_{dr} denote electricity and driver costs, e_t is electricity consumed in the time step t, Δt is the time step duration. The penalties corresponds to the average costs incurred during hazardous situations. The reward R_{tar} denotes the revenue that can be achieved by the truck when it reaches the target.

  3. **Vector-based reward for multi-objective learning:** In subsequent studies, the problem was addressed using a multi-objective RL formulation where the reward function is vector-valued, given by,

  $$\mathbf{r}_t = \begin{bmatrix} I_{tar} R_{tar} - I_c P_c \\ -C_{dr} \Delta t \\ -C_{el} e_t \end{bmatrix} \tag{2.4}$$

  Each component represents a specific objective: safety, quantified in terms of collisions and successful completion; energy efficiency,

quantified by energy cost; and time efficiency, quantified by driver cost. Vector-based reward formulation is described in more detail later in this chapter and is used in Paper II.

These reward formulations reflect the evolution of the problem formulation in this thesis, from a basic safety-focused design to a more realistic economic perspective and finally to a multi-objective framework suitable for heavy-duty highway driving.

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

2.3 Reinforcement Learning

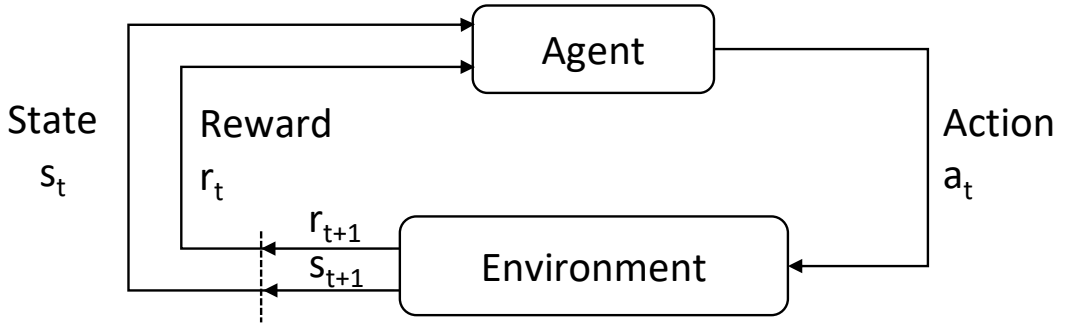

Figure 2.1: Schematic diagram of interaction between agent and environment in Reinforcement Learning.

Reinforcement Learning (RL) is a machine learning paradigm that learns optimal sequential decision making policies through interaction with an environment [20]. An RL problem is commonly modeled as MDP, defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ as described in section 2.1. At each time step t, an agent observes the current state of the environment $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}$ according to a policy $\pi(a \mid s)$. As a consequence of this action, the agent receives a scalar reward r_{t+1} as feedback, and the environment transitions to a new state s_{t+1}. Through repeated interaction with the environment, the agent aims to learn a policy that maximizes the discounted cumulative return G_t given by,

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.5}$$

The expected discounted return obtained by starting from state s, and following policy π is captured by *state-value* function defined as,

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[G_t \mid s_t = s\right] \tag{2.6}$$

Similarly, *action-value* $Q^{\pi}(s, a)$ represents the expected return obtained by taking action a at state s and subsequently following policy π,

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}\left[G_t \mid s_t = s, a_t = a\right] \tag{2.7}$$

An optimal policy π^* maximizes the expected return for all states and satisfies

$$V^*(s) = \max_\pi V^\pi(s), \quad Q^*(s,a) = \max_\pi Q^\pi(s,a). \tag{2.8}$$

for all $s \in \mathcal{S}$ and all $a \in \mathcal{A}$.

When the state or action spaces are large or continuous, function approximation is commonly used to represent value functions or policies. Deep reinforcement learning leverages neural networks as function approximators, enabling RL methods to scale to high-dimensional settings.

Reinforcement learning algorithms can be broadly categorized into *value-based* methods and *policy-based* methods. Value-based methods aim to learn an approximation of the action-value function $Q(s,a)$ from which a policy is implicitly derived by selecting the action that maximizes the estimated value. In contrast, policy-based methods directly parameterize the policy $\pi_\theta(a \mid s)$ and optimize it by maximizing the expected return using optimization methods such as gradient ascent.

From the perspective of environment modeling, RL methods are commonly divided into *model-based* and *model-free* approaches. Model-based methods explicitly learn or assume access to a model of the transition dynamics and use that model for planning the actions. In contrast, model-free methods do not assume the transition dynamics to be known but instead learn decision making strategies directly through the interactions with the environment. Model-free RL is particularly attractive in complex systems where accurate modeling of the environment is difficult, such as the tactical decision making problem.

In this thesis, we have considered the following model-free reinforcement learning methods, including both value-based and policy-based algorithms.

2.3.1 Deep Q-Network (DQN)

DQN is a value-based reinforcement learning algorithm that approximate the action-value function $Q(s,a)$ using a deep neural network [21]. The objective is to learn an approximation of the optimal Q-function from which a policy is implicitly derived by selecting the action that maximizes the estimated Q-value at each state. The optimal Q-function that satisfies the *Bellman optimality equation*,

$$Q^*(s,a) = \mathbb{E}_{s'}\left[r + \gamma \max_{a'} Q^*(s',a')\right]. \tag{2.9}$$

which expresses the recursive relationship between the value of a state–action pair and the expected return obtained from the subsequent state. In practice, this equation is used as a target for learning.

Training stability is a key challenge when combining Q-learning with non-linear function approximation. To address this, DQN introduces two important mechanisms: experience replay and a target network. Experience replay stores previously observed transitions in a replay buffer and samples mini-batches uniformly during training, which reduces temporal correlations between consecutive updates. The target network, whose parameters are updated less frequently than those of the main network, is used to compute stable target values, and it is periodically synchronized with the main network.

DQN employs temporal-difference (TD) learning, where value estimates are updated by minimizing the difference between the current Q-value prediction and a bootstrap target that combines the observed immediate reward with an estimate of future returns [20]. The main network parameterized by θ is trained by minimizing the temporal-difference loss,

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta)\right)^2\right], \tag{2.10}$$

where $\mathcal{D}$ is the experience replay buffer and θ^- denotes the parameters of a target network. By iteratively minimizing this loss, the Q-network converges toward an approximation of the optimal action-value function.

2.3.2 Advantage Actor-Critic (A2C)

A2C is a policy-based reinforcement learning algorithm that belongs to the family of actor–critic methods [22]. Actor–critic approaches combine the strengths of value-based and policy-based methods by learning a parameterized policy (the actor) together with a value function (the critic), which provides a learning signal to guide policy updates.

In A2C, the actor represents a stochastic policy $\pi_\theta(a \mid s)$, while the critic estimates the state-value function $V_\phi(s)$, which approximates the expected return from a given state under the current policy. The critic is used to reduce the variance of the policy gradient estimates by evaluating how favorable an action is relative to the expected value of the state.

Policy optimization in A2C is based on the advantage function, defined as

$$A_t = r_{t+1} + \gamma V_\phi(s_{t+1}) - V_\phi(s_t), \tag{2.11}$$

which measures how much better or worse the observed result of taking an action in a state is compared to the current value estimate for that state. A positive advantage indicates that the action taken performed better than expected, while a negative advantage suggests the opposite. This advantage is used to update the actor's policy parameters θ as

$$\Delta\theta = \alpha_\theta \, A_t \, \nabla_\theta \log \pi_\theta(a_t \mid s_t), \tag{2.12}$$

where α_θ is the actor's learning rate. The critic's parameters ϕ are updated to minimize the difference between its value estimate and the observed return,

$$\Delta\phi = \alpha_\phi \, A_t \, \nabla_\phi V_\phi(s_t), \tag{2.13}$$

with α_ϕ being the critic's learning rate. By iteratively updating both actor and critic in this manner, the algorithm converges toward a policy that maximizes expected return.

2.3.3 Proximal Policy Optimization (PPO)

PPO is an actor-critic algorithm designed to improve the stability and reliability of policy gradient methods [23]. Compared to other policy gradient approaches,

PPO constrains the magnitude of policy updates to avoid large changes that degrade performance.

PPO optimizes a clipped surrogate objective function based on the probability ratio between the new and old policies,

$$\rho_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \tag{2.14}$$

The clipped surrogate objective is defined as

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}\left[\min\left(\rho_t(\theta)A_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right], \tag{2.15}$$

where A_t is the advantage function and ϵ is a hyperparameter that controls the allowed deviation from the previous policy. The clipping operation ensures that the policy update does not move too far in a single step, stabilizing learning while still allowing improvement in the expected return. An entropy term is often added to the actor loss to encourage exploration. The critic in PPO is trained similarly to A2C by minimizing the squared error between its value estimates and observed returns.

PPO strikes a balance between sample efficiency and training stability, making it well-suited for complex environments with high-dimensional state and action spaces.

2.4 Multi-Objective Reinforcement Learning

The reinforcement learning formulation introduced in Section 2.3 assumes a scalar reward signal and a single notion of optimality. However, many real-world decision making problems, including the tactical decision making problem considered in this thesis, involve multiple, often conflicting objectives such as safety, efficiency, and traffic flow. To explicitly model and reason about trade-offs between such objectives, these problems can be formulated within a multi-objective reinforcement learning (MORL) framework.

In MORL, the standard Markov Decision Process is extended to a Multi-Objective Markov Decision Process (MOMDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \mathbf{R}, \gamma)$ where the state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics P, and discount factor γ retain the same interpretation as in the single-objective MDP. The key distinction lies in the reward function $\mathbf{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, which provides a vector-valued reward $\mathbf{r}_{t+1} = \mathbf{R}(s_t, a_t, s_{t+1})$ at each timestep t. Here, $d \in \mathbb{N}$ denotes the number of objectives, and each component of the reward vector corresponds to a distinct objective.

As in a single-objective setting, a policy $\pi(a \mid s)$ defines the agent's decision making behavior. The value of a policy is naturally generalized to a vector-valued state-value function, defined as the expected discounted cumulative reward for each objective,

$$\mathbf{V}^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \mathbf{r}_{t+k+1} \;\middle|\; s_t = s\right], \tag{2.16}$$

where $\mathbf{V}^{\pi}(s) \in \mathbb{R}^d$. Unlike in single-objective RL, these value vectors cannot, in general, be totally ordered. Optimality in MORL is therefore defined in terms of *Pareto dominance*. A policy π' is said to dominate another policy π if it performs at least as well in all objectives and strictly better in at least one. The set of non-dominated policies forms the *Pareto front*, as shown in Figure 2.2, which characterizes the achievable trade-offs among objectives. Consequently, the goal in MORL is not to learn a single optimal policy, but rather to approximate this Pareto front.

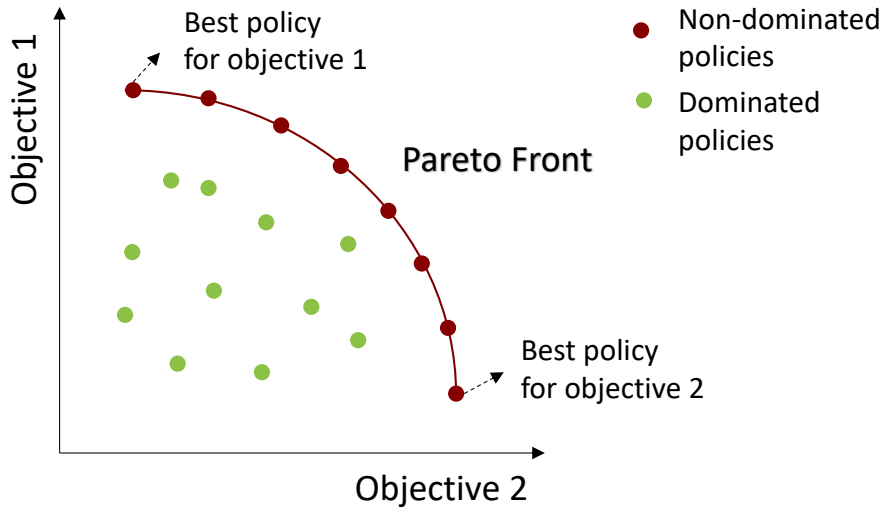

Figure 2.2: Pareto front for a multi-objective optimization problem with two objectives.

To achieve this, the agent must explore the policy space to identify policies corresponding to different objective trade-offs, and thereby approximate the set of Pareto-optimal solutions. To enable the use of standard reinforcement learning methods for policy optimization, scalarization is introduced via a utility function

$$u : \mathbb{R}^d \to \mathbb{R},$$

which maps the multi-objective value vector to a scalar value. The most widely used scalarization, and the one adopted in this thesis, is linear scalarization,

$$u(\mathbf{V}^{\pi}; \mathbf{w}) = \mathbf{w}^{\top}\mathbf{V}^{\pi} = \sum_{i=1}^{d} w_i V_i^{\pi}, \tag{2.17}$$

where the weight (preference) vector $\mathbf{w} \in \mathbb{R}^d$ satisfies $w_i \geq 0$ for all i and $\sum_{i=1}^{d} w_i = 1$. Each weight vector defines a corresponding single-objective optimization problem with scalarized rewards

$$r_{\mathbf{w}}(s, a, s') = \mathbf{w}^{\top}\mathbf{r}(s, a, s'). \tag{2.18}$$

The set of policies that are optimal for some weight vector $\mathbf{w}$ constitutes the Convex Hull (CH) of the achievable value set. The minimal subset containing

one optimal policy per weight vector is known as the Convex Coverage Set (CCS), which provides a compact representation of all linearly Pareto-optimal solutions.

In practice, approximating CCS by solving a separate reinforcement learning problem for every possible weight vector $\mathbf{w}$ is computationally infeasible. Consequently, several algorithmic frameworks have been proposed to efficiently approximate the Pareto front or CCS, typically by employing extended versions of traditional RL algorithms for policy optimization. These include evolutionary approaches that maintain a population of policies to evolve toward the Pareto optimal set [24]; preference-conditioned policies, such as Pareto Conditioned Networks, that use a neural network to generalize across desired returns or weights [25]; and decomposition-based methods that break the Pareto front into simpler sub-problems solved concurrently [26]. In Paper II, appended to this thesis, we implement a multi-objective extension of PPO and utilize a sample-efficient approach for prioritizing the weight vectors for training based on Generalized Policy Improvement (GPI) [27], [28]. This strategy identifies weight vectors where the existing learned policies have the highest potential for improvement, allowing for a more effective refinement of the CCS.

2.5 Traffic Environment Modeling

To enable systematic and reproducible investigation of tactical decision making for autonomous heavy-duty trucks, a scalable simulation environment is developed using the *Simulation of Urban Mobility (SUMO)* platform [29]. SUMO is an open-source microscopic traffic simulator that models individual vehicles and their interactions while supporting large-scale experiments with heterogeneous traffic compositions.

The simulation environment represents a highway driving scenario as shown in Figure 2.3, composed of a three-lane road segment with configurable length. Traffic consists of a heterogeneous mixture of passenger vehicles and heavy-duty trucks. The ego heavy-duty truck is controlled by a learning-based tactical decision making policy while surrounding vehicles are controlled using SUMO's default car-following and lane-changing models. A moving window of surrounding vehicles is maintained around the ego truck to maintain a stationary traffic distribution, with vehicles dynamically re-spawned at the boundaries of the window. Traffic density, vehicle composition, and their characteristics are configurable parameters that enable experiments in different levels of congestion and traffic settings. Simulation episodes end when the ego vehicle reaches the end of the road segment, a crash occurs, or a predefined maximum simulation time is exceeded.

2.5.1 Car-Following Models

Car-following models describe how a vehicle adjusts its speed based on the motion of the leading vehicle. These models aim to maintain safe distances with leading vehicle and avoid collisions while not lag too far behind.

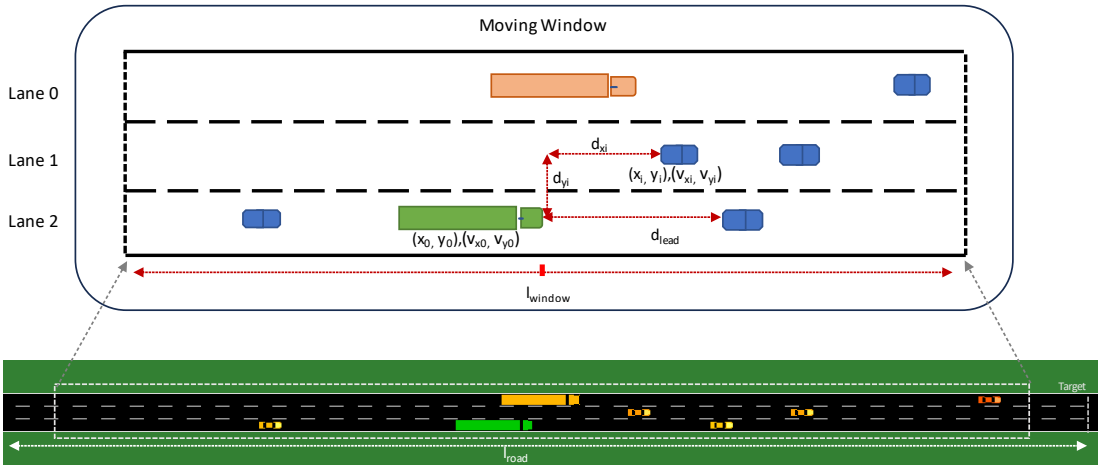

Figure 2.3: Simulated traffic environment in SUMO with the illustration of moving window. The ego vehicle is described by its position and velocity $(x_0, y_0, v_{x0}, v_{y0})$ and surrounding vehicles are characterized by $(x_i, y_i, v_{xi}, v_{yi})$. The relative longitudinal and lateral distances of surrounding vehicles are represented by d_{xi}, d_{yi} while d_{lead} denotes the longitudinal distance to the leading vehicle. The moving window size l_{window}, road length l_{road} and traffic density are configurable parameters.

There are several car-following models in traffic literature. In this work, we implement Intelligent Driver Model (IDM) [30] for low-level speed control for the ego truck. IDM computes the longitudinal acceleration a of following vehicle α as a function of its current speed v_α, the speed difference Δv_α with respect to the leading vehicle, and the gap s_α to the vehicle ahead:

$$\dot{v}_\alpha = \frac{\mathrm{d}v_\alpha}{\mathrm{d}t} = a\left(1 - \left(\frac{v_\alpha}{v_0}\right)^\delta - \left(\frac{s^*\left(v_\alpha, \Delta v_\alpha\right)}{s_\alpha}\right)^2\right) \tag{2.19}$$

where the desired gap $s^*(v, \Delta v)$ is defined as

$$s^*\left(v_\alpha, \Delta v_\alpha\right) = s_0 + v_\alpha T + \frac{v_\alpha \Delta v_\alpha}{2\sqrt{ab}} \tag{2.20}$$

Here, v_0 is the desired speed of the vehicle, s_0 is the minimum spacing, T is the desired time gap to the leading vehicle, a is the maximum acceleration, b is the comfortable deceleration and δ is an exponent (typically 4).

The acceleration computed by IDM is composed of two terms. The first term, $a\left(1 - \left(\frac{v_\alpha}{v_0}\right)^\delta\right)$ defines the desired acceleration on a free road, driving the vehicle towards the desired speed v_0. As the current speed v_α approaches v_0, this term smoothly decreases from the maximum acceleration a to zero. The second term $-a\left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{s}\right)^2$ represents the deceleration induced by the leading vehicle, comparing the actual gap s_α to the desired gap $s^*(v_\alpha, \Delta v_\alpha)$. This term ensures that the vehicle decelerates smoothly when approaching a slower vehicle, maintaining a safe following distance and avoiding collisions.

The desired gap accounts for minimum spacing, desired time gap, and relative speed, allowing IDM to produce realistic transitions between free driving, car-following, and braking.

The surrounding vehicles in the simulation uses the default car-following model implemented in SUMO, known as Krauss model [31]. Unlike IDM, which computes acceleration, the Krauss model determines the maximum safe velocity for a following vehicle by considering the current speed of the leading vehicle, the gap between vehicles, the driver's reaction time, and the braking capabilities of both vehicles. The model also incorporates stochastic noise to capture variability in driver behavior, ensuring realistic traffic dynamics.

2.5.2 Lane-Change Models

Lane-change models describe how vehicles decide when and how to change lanes in multi-lane traffic. The default lane-change model in SUMO, known as LC2013 [32], evaluates both motivations and safety criteria for performing a lane change. Motivations include *strategic* lane changes to follow the planned route, *cooperative* lane changes to allow other vehicles to perform maneuvers, *tactical* lane changes to gain speed or improve travel time, and lane changes due to *obligations*, such as clearing the overtaking lane. The safety criterion ensures that sufficient gaps exist in the target lane to avoid collisions. Once a lane-change decision is made, LC2013 computes the lateral trajectory required for the vehicle to transition smoothly to the target lane.

In this work, the lane changes for surrounding vehicles are completely controlled by LC2013 model. However, for the ego vehicle, lane changes are initiated by a learned policy whereas the LC2013 is used only to compute the lateral motion for performing the lane change.

Chapter 3

Summary of Included Papers

In this chapter, we provide a summary of the two papers appended to this thesis.

3.1 Paper I

The first paper included in this thesis addresses tactical decision making for autonomous heavy-duty trucks in highway traffic, focusing on Adaptive Cruise Control (ACC) and lane change maneuvers. The study uses a simulation of a three-lane, 3 km highway segment with 15 vehicles. The ego vehicle is modeled as a truck-trailer combination and all surrounding vehicles are passenger cars.

Many existing studies apply RL directly to safety-critical decisions such as speed control. However, RL policies inherently involve uncertainties that pose risks in safety-critical applications. To mitigate this, the paper proposes a hierarchical decision making framework that integrates RL with physics-based controllers. High-level decisions including changing desired speed, desired timegap or lane are handled by RL, while low-level control is executed by traditional controllers. Specifically, we use the Intelligent Driver Model (IDM) for low-level speed regulation and SUMO's default LC2013 model handles the execution of lane changes. This hierarchical architecture is evaluated against a baseline that includes only an RL agent making decisions about both speed and lane changes. Using DQN, A2C, and PPO algorithms with a safety-focused reward function, the results show that the hierarchical approach consistently outperforms the baseline, regardless of the RL algorithm used.

For heavy-duty trucks, operational efficiency and cost are critical due to their commercial nature. This motivates a framework that explicitly incorporates economic considerations alongside safety objectives. To capture these, we design a reward function based on the Total Cost of Operation (TCOP), which jointly reflects efficiency, safety, and operational cost within a single learning objective. These composite rewards also introduce challenges related

to conflicting objectives and learning stability.

The paper systematically explores strategies for handling multi-component rewards within a scalar RL framework including weighted scalarization of reward components, reward normalization and reward based curriculum learning. Weighted scalarization combines reward components with fixed weights, offering simplicity and efficiency but requiring careful tuning. Reward normalization is implemented by dividing each cost component by the distance traveled per step, improving stability and convergence. Curriculum Reinforcement Learning (CRL) is explored as a strategy to manage complexity in multi-component rewards. By gradually introducing new reward aspects, the agent first learns stable behaviors for simpler objectives before facing complex trade-offs. 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. In general, this complex cost-based reward function poses challenges in terms of computation and stability, compared to a simple safety-focused reward function.

Contributions Deepthi Pathare performed the main work, including the design and implementation of the framework and simulation environment, training of the models, and conducting the experimental studies. The project was jointly supervised by Morteza Haghir Chehreghani and Leo Laine.

3.2 Paper II

Building on the insights and limitations identified in the first paper, the second paper shifts focus from scalarized reward formulations to an explicit treatment of competing objectives in tactical decision making for heavy-duty trucks. While the earlier work demonstrates that the tactical decision making problem can be tackled using a deep RL framework with a scalar reward function, it also reveals challenges related to reward tuning and convergence. These observations motivate a reformulation of the decision making problem that directly acknowledges its multi-objective nature rather than collapsing multiple components into a single scalar signal.

To address this, the second paper formulates the tactical decision making problem within a multi-objective reinforcement learning (MORL) framework. Instead of aggregating different aspects of driving performance into a single reward signal, MORL models each objective separately using a vector of rewards. In this work, we consider three objectives: safety, quantified in terms of collisions and successful task completion; energy efficiency, quantified through energy cost; and time efficiency, quantified through driver cost.

In contrast to the first paper, which focuses on learning a single policy optimized for multiple objectives, this work aims to learn a set of Pareto-optimal policies. In the multi-objective setting, a policy is considered Pareto-optimal if no objective can be further improved without degrading at least one other objective. By approximating the Pareto frontier, the proposed framework provides a structured representation of the achievable trade-offs

between safety, energy consumption, and travel time. This shifts the emphasis from identifying a single best policy to exploring a space of viable tactical behaviors. This approach better reflects real-world freight transport operations, where improvements along one dimension often incur costs in another, and where no single policy is universally optimal across all operational preferences.

We use Proximal Policy Optimization (PPO) as the underlying reinforcement learning algorithm due to its proven performance in Paper I. We develop a multi-objective extension of PPO architecture that employs a vector-valued critic together with an actor network producing per-objective action logits, while applying scalarization only at the loss level. The actor–critic networks are trained iteratively by selecting preference weight vectors expected to yield the highest improvement based on Generalized Policy Improvement (GPI).

The experimental results demonstrate that the proposed MORL framework successfully learns diverse and interpretable tactical driving behaviors corresponding to different trade-offs between successful completion, driver cost and energy cost. The learned policies exhibit consistent and meaningful structure across the Pareto frontier, revealing clear behavioral patterns such as conservative, energy efficient driving versus more time efficient but aggressive strategies. Furthermore, experiments across different traffic density settings show that the nature of the learned Pareto-optimal policies depends on the interaction dynamics of the environment. While the MORL framework captures trade-offs between objectives, the specific behaviors that emerge along the Pareto frontier vary with traffic conditions, reflecting differences in vehicle interactions and complexity.

Contributions Deepthi Pathare performed the main work, including the design and implementation of the MORL framework, training of the models, and conducting the experimental studies. The project was jointly supervised by Morteza Haghir Chehreghani and Leo Laine.

Chapter 4

Concluding Remarks and Future Work

In this thesis, we investigate learning-based tactical decision making for autonomous heavy-duty trucks, focusing on high-level driving decisions such as Adaptive Cruise Control and lane selection. The two papers in the thesis provide a coherent methodological progression, examining both the capabilities and limitations of reinforcement learning for economically meaningful autonomous driving.

The first paper demonstrates that reinforcement learning can be effectively applied to tactical decision making when combined with a hierarchical control architecture. By limiting RL to perform high-level tactical decisions and delegating low-level tactical control to physics-based controllers, the approach improves safety and overall performance. In this study, we also introduce a realistic reward function that captures the economic aspects of truck driving. However, the findings highlight the challenges of scalar reward formulations in such settings in terms of convergence and learning stability.

The second paper addresses these limitations through a multi-objective reinforcement learning formulation, explicitly modeling safety, energy efficiency, and time efficiency as separate objectives. This approach approximates a Pareto front, providing transparent insight into trade-offs, enabling flexible choices of policies without retraining.

While the presented framework demonstrates the feasibility and benefits of learning-based tactical decision making for autonomous heavy-duty trucks, several important extensions remain for future research. First, the current hierarchical framework adopts a simplified action space in which the agent selects only one discrete maneuver at a time, such as performing Adaptive Cruise Control adjustments or initiating a lane change. However, driving in the real-world requires the simultaneous coordination of longitudinal and lateral actions. Future work will therefore extend the action space to enable concurrent longitudinal and lateral decision making. In addition, mechanisms will be incorporated to facilitate cooperative interactions with surrounding vehicles, particularly during lane change maneuvers. Such extensions are

expected to enhance behavioral realism, promote cooperative driving behavior, and enable smoother maneuver execution.

Second, while the current work focuses primarily on truck-level objectives such as safety, efficiency, and operational costs, broader system-level objectives remain largely unexplored. Heavy-duty trucks play a significant role in overall traffic dynamics, and their decisions can substantially influence traffic flow efficiency and congestion patterns. Future work will therefore investigate the integration of macro-level goals, such as traffic flow efficiency and cooperative behavior, into the decision making framework. To achieve more realistic and interactive traffic modeling, surrounding vehicles will be represented using multi-agent reinforcement learning formulations rather than predefined behavioral models. The framework should also be further tuned and evaluated in more complex scenarios to fully assess robustness and scalability.

Finally, integrating real-world operational data collected from trucks through offline reinforcement learning techniques would enable more informed decision making, improve generalization, and support data-driven calibration of objective trade-offs grounded in operational realities.

Bibliography

[1] J. Nowakowska-Grunt and M. Strzelczyk, "The current situation and the directions of changes in road freight transport in the european union," *Transportation Research Procedia*, vol. 39, pp. 350–359, 2019, 3rd International Conference "Green Cities – Green Logistics for Greener Cities", Szczecin, 13-14 September 2018, ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2019.06.037. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146519301255 (cit. on p. 3).

[2] H. J. Walnum and M. Simonsen, "Does driving behavior matter? an analysis of fuel consumption data from heavy-duty trucks," *Transportation Research Part D: Transport and Environment*, vol. 36, pp. 107–120, 2015, ISSN: 1361-9209. DOI: https://doi.org/10.1016/j.trd.2015.02.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920915000255 (cit. on p. 3).

[3] D. C. Quiros, J. Smith, A. Thiruvengadam, T. Huai and S. Hu, "Greenhouse gas emissions from heavy-duty natural gas, hybrid, and conventional diesel on-road trucks during freight transport," *Atmospheric Environment*, vol. 168, pp. 36–45, 2017, ISSN: 1352-2310. DOI: https://doi.org/10.1016/j.atmosenv.2017.08.066. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231017305794 (cit. on p. 3).

[4] H. Zubaidi, A. Alnedawi, I. Obaid and M. G. Abadi, "Injury severities from heavy vehicle accidents: An exploratory empirical analysis," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 9, no. 6, pp. 991–1002, 2022, ISSN: 2095-7564. DOI: https://doi.org/10.1016/j.jtte.2021.02.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095756422000988 (cit. on p. 3).

[5] R. Schindler, M. Jänsch, A. Bálint and H. Johannsen, "Exploring european heavy goods vehicle crashes using a three-level analysis of crash data," *International Journal of Environmental Research and Public Health*, vol. 19, no. 2, 2022, ISSN: 1660-4601. DOI: 10.3390/ijerph19020663. [Online]. Available: https://www.mdpi.com/1660-4601/19/2/663 (cit. on p. 3).

[6] E. Papadimitriou, C. Schneider, J. Aguinaga Tello, W. Damen, M. Lomba Vrouenraets and A. ten Broeke, "Transport safety and human factors in the era of automation: What can transport modes learn from each other?" *Accident Analysis Prevention*, vol. 144, p. 105 656, 2020, ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2020.105656. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457520305558 (cit. on p. 3).

[7] W. Schwarting, J. Alonso-Mora and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. Volume 1, 2018, pp. 187–210, 2018, ISSN: 2573-5144. DOI: https://doi.org/10.1146/annurev-control-060117-105157. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-control-060117-105157 (cit. on p. 3).

[8] B. Paden, M. Čáp, S. Z. Yong, D. Yershov and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Transactions on intelligent vehicles*, vol. 1, no. 1, pp. 33–55, 2016 (cit. on p. 3).

[9] P. Nilsson, L. Laine, N. van Duijkeren and B. Jacobson, "Automated highway lane changes of long vehicle combinations: A specific comparison between driver model based control and non-linear model predictive control," in *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 2015, pp. 1–8. DOI: 10.1109/INISTA.2015.7276790 (cit. on p. 3).

[10] J. Suh, H. Chae and K. Yi, "Stochastic model-predictive control for lane change decision of automated driving vehicles," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 4771–4782, 2018. DOI: 10.1109/TVT.2018.2804891 (cit. on p. 3).

[11] S. Yu, M. Hirche, Y. Huang, H. Chen and F. Allgöwer, "Model predictive control for autonomous ground vehicles: A review," *Autonomous Intelligent Systems*, vol. 1, no. 1, p. 4, 2021 (cit. on p. 3).

[12] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2022. DOI: 10.1109/TITS.2021.3054625 (cit. on p. 4).

[13] J. Zhao, Y. Wu, R. Deng, S. Xu, J. Gao and A. Burke, "A survey of autonomous driving from a deep learning perspective," *ACM Comput. Surv.*, vol. 57, no. 10, May 2025, ISSN: 0360-0300. DOI: 10.1145/3729420. [Online]. Available: https://doi.org/10.1145/3729420 (cit. on p. 4).

[14] Z. Tian et al., "Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey," *arXiv preprint arXiv:2501.01886*, 2025 (cit. on p. 4).

[15] S. Chen et al., "Toward the robustness of autonomous vehicles in the ai era," *The Innovation*, vol. 6, no. 3, 2025 (cit. on p. 4).

[16] Z. Zhang, E. Demir, R. Mason and C. Di Cairano-Gilfedder, "Understanding freight drivers' behavior and the impact on vehicles' fuel consumption and co2e emissions," *Operational Research*, vol. 23, no. 4, p. 59, 2023 (cit. on p. 4).

[17] J. Lee, T. Oh and J. Yoo, "Adaptive longitudinal speed control for heavy-duty vehicles considering actuator constraints and disturbances using simulation validation," *Applied Sciences*, vol. 15, no. 13, 2025, ISSN: 2076-3417. DOI: 10.3390/app15137327. [Online]. Available: https://www.mdpi.com/2076-3417/15/13/7327 (cit. on p. 4).

[18] L. Hashimy, I. Castillo, W. Schildorfer and M. Neubauer, "Autonomous heavy-duty vehicles in logistics: Market trends, opportunities, and barriers," in *Transport Transitions: Advancing Sustainable and Inclusive Mobility*, C. McNally, P. Carroll, B. Martinez-Pastor, B. Ghosh, M. Efthymiou and N. Valantasis-Kanellos, Eds., Cham: Springer Nature Switzerland, 2026, pp. 585–592, ISBN: 978-3-032-06763-0 (cit. on p. 4).

[19] J. Engström et al., "Deployment of automated trucking: Challenges and opportunities," *Road Vehicle Automation 5*, pp. 149–162, 2018 (cit. on p. 4).

[20] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Second. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html (cit. on pp. 11, 13).

[21] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, ISSN: 00280836. [Online]. Available: http://dx.doi.org/10.1038/nature14236 (cit. on p. 12).

[22] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan and K. Q. Weinberger, Eds., ser. Proceedings of Machine Learning Research, vol. 48, New York, New York, USA: PMLR, 2016, pp. 1928–1937. [Online]. Available: https://proceedings.mlr.press/v48/mniha16.html (cit. on p. 13).

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, *Proximal policy optimization algorithms*, 2017. arXiv: 1707.06347 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1707.06347 (cit. on p. 13).

[24] O. S. Ajani, I. Fenyom, D. Darlan and R. Mallipeddi, "Prediction-guided multi-objective reinforcement learning with corner solution search," *Computers and Electrical Engineering*, vol. 122, p. 109 964, 2025 (cit. on p. 16).

[25] M. Reymond, E. Bargiacchi and A. Nowé, "Pareto conditioned networks," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '22, Virtual Event, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2022, 1110–1118, ISBN: 9781450392136 (cit. on p. 16).

[26] F. Felten, E.-G. Talbi and G. Danoy, "Multi-objective reinforcement learning based on decomposition: A taxonomy and framework," *Journal of Artificial Intelligence Research*, vol. 79, pp. 679–723, 2024 (cit. on p. 16).

[27] A. Barreto, S. Hou, D. Borsa, D. Silver and D. Precup, "Fast reinforcement learning with generalized policy updates," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 079–30 087, 2020. DOI: 10.1073/pnas.1907370117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1907370117. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1907370117 (cit. on p. 16).

[28] L. N. Alegre, A. L. C. Bazzan, D. M. Roijers, A. Nowé and B. C. da Silva, "Sample-efficient multi-objective learning via generalized policy improvement prioritization," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '23, London, United Kingdom: International Foundation for Autonomous Agents and Multiagent Systems, 2023, 2003–2012, ISBN: 9781450394321 (cit. on p. 16).

[29] P. A. Lopez et al., "Microscopic traffic simulation using sumo," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, HI, USA: IEEE Press, 2018, 2575–2582, ISBN: 978-1-7281-0321-1. DOI: 10.1109/ITSC.2018.8569938. [Online]. Available: https://doi.org/10.1109/ITSC.2018.8569938 (cit. on p. 16).

[30] M. Treiber, A. Hennecke and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000 (cit. on p. 17).

[31] S. Krauss, P. Wagner and C. Gawron, "Metastable states in a microscopic model of traffic flow," *Phys. Rev. E*, vol. 55, 5 1997 (cit. on p. 18).

[32] J. Erdmann, "Lane-changing model in sumo," in *Proceedings of the SUMO2014 Modeling Mobility with Open Data*, ser. Reports of the DLR-Institute of Transportation SystemsProceedings, vol. 24, 2014 (cit. on p. 18).