IEEE Open Journal of
Intelligent Transportation
Systems

Received 26 June 2020; revised 27 August 2020; accepted 25 September 2020. Date of publication 29 September 2020; date of current version 5 November 2020.

Digital Object Identifier 10.1109/OJITS.2020.3027518

Optimised Traffic Light Management Through Reinforcement Learning: Traffic State Agnostic Agent vs. Holistic Agent With Current V2I Traffic State Knowledge

JOHANNES V. S. BUSCH [1], VINCENT LATZKO [1] (Graduate Student Member, IEEE), MARTIN REISSLEIN [2] (Fellow, IEEE), AND FRANK H. P. FITZEK [1] (Senior Member, IEEE)

[1]Deutsche Telekom Chair of Communication Networks, Technical University of Dresden, 01069 Dresden, Germany

[2]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA

CORRESPONDING AUTHOR: M. REISSLEIN (e-mail: reisslein@asu.edu)

This work was supported in part by the Federal Ministry of Education and Research of the Federal Republic of Germany (BMBF) in the framework of the project AI4Mobile under Grant 16KIS1178, and in part by the project European Union H2020–ECSEL–AutoDrive under Grant 737469.

ABSTRACT Traffic light control falls into two main categories: *Agnostic* systems that do not exploit knowledge of the current traffic state, e.g., the positions and velocities of vehicles approaching intersections, and *holistic* systems that exploit knowledge of the current traffic state. Emerging fifth generation (5G) wireless networks enable Vehicle-to-Infrastructure (V2I) communication to reliably and quickly collect the current traffic state. However, to the best of our knowledge, the optimized traffic light management without and with current traffic state information has not been compared in detail. This study fills this gap in the literature by designing representative Deep Reinforcement Learning (DRL) agents that learn the control of *multiple* traffic lights without and with current traffic state information. Our agnostic agent considers mainly the current phase of all traffic lights and the expired times since the last change. In addition, our holistic agent considers the positions and velocities of the vehicles approaching the intersections. We compare the agnostic and holistic agents for simulated traffic scenarios, including a road network from Barcelona, Spain. We find that the holistic system substantially increases average vehicle velocities and flow rates, while reducing CO_2 emissions, average wait and trip times, as well as a driver stress metric.

INDEX TERMS Deep reinforcement learning (DRL), intelligent transportation system (ITS), intersection control, vehicle-to-infrastructure communication (V2I).

I. INTRODUCTION

A. MOTIVATION

EFFECTIVE transportation systems are a key requirement for economic competitiveness and environmental sustainability. Inefficiencies in traffic regulation lead to congestion, causing high costs and commuter delays. In the European Union (EU), the cost of congestion was estimated to amount to 1% of the annual GDP in 2017 [1]. In 2019, the cost of congestion in the U.S. amounted to 88 Billion Dollars (0.41% of GDP) [2]. Especially in dense metropolises, commuters annually spend up to 200 hours stuck in traffic [2]. In addition to the economic burden, increased emissions due to congestion have undesirable environmental and social repercussions [3], [4]. Mitigating congestion is of paramount importance to transportation authorities and different solution strategies have been proposed, many requiring expensive and time-consuming construction work on the road network. Especially in dense city centers, these measures are often obstructed by existing infrastructures. Therefore, a compelling approach is the more efficient control of traffic through intelligent traffic light systems [5]–[9], which requires no expansion of the road infrastructure and is comparably cheap and easy to implement.

The review of this article was arranged by Associate Editor Jiaqi Ma.

In recent years, fast and reliable wireless technology has given rise to an increased interest in so-called Vehicle-to-Infrastructure (V2I) communication and its applications. Current and upcoming standards, such as IEEE 802.11p, LTE-V, and 5G, allow the exchange of information between individual vehicles and the traffic infrastructure, eventually providing the infrastructure with holistic knowledge of the current state of the traffic system. This should, in theory, enable highly informed control decisions and facilitate congestion mitigation. However, traditional traffic control paradigms are unfit to leverage the dense stream of state information towards making better control decisions. Novel algorithms are therefore required to cope with the staggering complexity of traffic control, under consideration of detailed state information.

Throughout the last decade, increasing computational capabilities and large datasets enabled the effective training of Deep Neural Networks (DNNs) and led to the advent of the field of Machine Learning (ML). Some of the most impressive ML achievements stem from the subfield of Reinforcement Learning (RL), that addresses complex control problems with learning-based approaches. In particular, the combination of Reinforcement Learning and Deep Neural Networks, referred to as Deep Reinforcement Learning (DRL), showed to be able to solve many intricate control problems, ranging from Atari arcade games [10] and the ancient board game of Go [11] to robotic control [12].

B. CONTRIBUTIONS AND STRUCTURE OF THIS ARTICLE

The main contribution of this study is to provide a detailed comparison of a representative state-of-the-art agnostic DRL agent that is oblivious to the current traffic state with a representative state-of-the-art holistic DRL agent that has knowledge of the current traffic state. We also compare a reward function that considers only the average vehicle velocity with a composite reward function that considers a weighted combination of the average vehicle velocity, vehicle flow rate, CO_2 emissions, and driver stress level. Importantly, we conduct these comparisons for road networks with multiple intersections, including a network with one main arterial road and several side roads, a network with a 3×3 grid of intersections, and a patch of the road network in Barcelona, Spain. The traffic state includes the positions and velocities of the vehicles approaching the various intersections as well as the numbers of vehicles on the various roads, which can be readily collected in real-time with 5G V2I communication.

We find that compared to the agnostic agent, the holistic agent achieves significantly higher average vehicle velocities and flow rates, as well as significantly shorter average trip times through the road networks. Moreover, the holistic agent significantly reduces the average wait times at traffic lights. Also, the holistic agent reduces the CO_2 emissions and a driver stress metric. Generally, the performance improvements with the holistic agent are more pronounced for relatively low traffic demands and for complex road networks. For high traffic demands at a single intersection, the holistic agent increases the average vehicle velocities only slightly compared to the agnostic agent. However, if there is low traffic at a single intersection, or multiple intersections are considered (for any traffic demand), then the holistic agent performs significantly better than the agnostic agent.

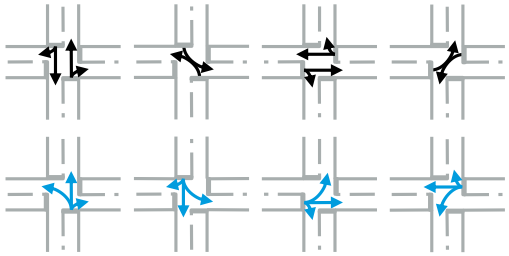


FIGURE 1. Two popular phase schemes. Each phase contains only compatible streams, which ensures safe operation.

This article is structured as follows. In Section II, we review the background of our problem. In Section III, we discuss related work. In Section IV, we introduce the designs of the representative agnostic and holistic DRL agents. In Section V, we describe the set-up of the simulation experiments and the detailed agnostic vs. holistic DRL agent comparison results. Finally, in Sections VI and VII, we summarize the obtained results and outline further research directions.

II. BACKGROUND

This section explains the relevant background of the related fields, namely, intersection control, V2I communication, and RL.

A. INTERSECTION CONTROL

Traditionally, traffic lights at an intersection sequentially go through different phases in which the right of way is granted to the different streams in a predefined sequence of individual phases, called the phase scheme [5]–[9]. A stream is defined as an allowed trajectory from an approach to an exit of the intersection [13]. Streams are called compatible if vehicles of two or more streams can cross the intersection without interfering; otherwise they are called antagonistic [14]. Fig. 1 shows two popular phase schemes, that both consist of only compatible streams.

Phase durations can either be computed in advance, using historical traffic statistics, or can be adapted according to the current traffic state that is measured through inductive loop sensors in the road pavement. As choosing the optimal phase times is not trivial and traffic systems are subject to tight real-time constraints, the utilization of the current traffic state often leads to suboptimal phase durations. Furthermore, in a dense traffic network, the signaling of one traffic light strongly affects the efficacy of the signaling of its surrounding traffic lights. For an optimal control

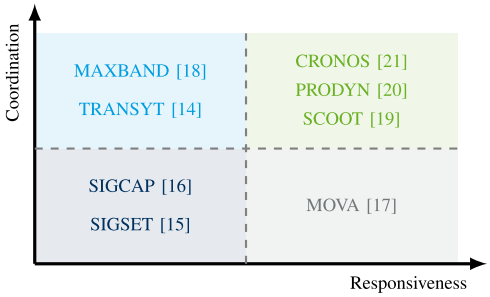


FIGURE 2. Popular traditional control strategies categorized along the coordination and the responsiveness axes.

policy, we consequently cannot merely optimize isolated intersections but need to jointly consider the parameters of multiple intersections to find a well-coordinated solution [14]. However, the complexity of the optimization problem increases exponentially with the number of considered traffic lights. Therefore, most existing traffic lights are optimized in isolation or in coordination with only a few nearby intersections. Traditional control methods can therefore roughly be categorized alongside two axes: *Responsiveness* to the current traffic situation and *coordination* between the individual intersections [14]. Fig. 2 shows some traditional control algorithms, categorized along the two axes. To ensure safe operation, phases cannot be changed arbitrarily. Rather, traffic lights have to undergo an amber period, which allows vehicles of those streams that lose the right of way to brake, and an all-red period, in that the intersection can be fully cleared as no further vehicles are allowed to enter. The lengths of the amber and the all-red periods depend on the speed limits of the approaches and the dimensions of the intersection and are typically not parameters of a traffic optimization algorithm [22].

B. VEHICLE TO INFRASTRUCTURE COMMUNICATION

Vehicle to Infrastructure (V2I) communication technologies enable a bidirectional exchange of information between individual vehicles and the traffic infrastructure [23]–[26]. The high safety-relevance and tight real-time constraints of traffic systems demand a reliable high-performance communication interface that provides very low latency and high throughput in conditions of high mobility and vehicular density. The IEEE 802.11p Wi-Fi standard [27] was introduced for local wireless access in Intelligent Transportation Systems [28]. IEEE 802.11p Wi-Fi enables data rates between 6 Mbps and 27 Mbps at a relatively short transmission range around 300 m [23]. Unfortunately, it suffers from scalability issues, unbounded delays, and lack of deterministic quality of service (QoS) guarantees [29].

Cellular technology provides an alternative that can overcome the limitations of the IEEE 802.11p Wi-Fi standard. In 2016, the Third Generation Partnership Project (3GPP) published the first version of Release 14, which features support for V2X communications [30]. This standard, referred to as LTE-V, offers increased reliability with respect to IEEE 802.11p. In the absence of a cellular connection (e.g., in rural areas), LTE-V can use the PC5 sidelink4 interface for direct device-to-device communication without the need for base stations. 3GPP Release 15 features V2X using the emerging 5G mobile Internet [30] which improves on LTE-V in every aspect. 5G is expected to achieve latencies as short as one millisecond and throughput of up 10 Gbit/s for up to 100 billion independent devices as well as greatly increased capacity [31]–[34]. 5G in combination with emerging low-latency multi-access edge computing [35]–[37] appears thus well suited to enable intricate real-time control decisions in V2X scenarios.

C. REINFORCEMENT LEARNING

The Reinforcement Learning (RL) framework deals with autonomous agents that navigate and explore their environment. The environments are framed as so-called Markov Decision Processes (MDPs). An agent's interactions with the environment are executed in discrete timesteps. In every timestep t, the agent observes the state s_t of its environment and takes an action a_t to influence its environment, resulting in a new state s_{t+1}. The mapping from states to actions is called the agent's policy $\pi(a_t|s_t)$, which may be deterministic or stochastic. The agent's behavior is evaluated by some numerical reward signal r_t. The agent thus strives to maximize a weighted sum of future rewards $g_t = \sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i+1}$, called the (discounted) return. The discount factor $\gamma \in [0, 1)$ defines how much the agent prefers rewards in the near future over those in the distant future.

Arguably the most popular RL algorithm is Q-Learning [38]. With Q-Learning, the agent estimates the return that it will obtain if it currently is in some particular state s, executes some action a, and always executes the action with the highest Q-value thereafter. This estimate is called the action-value or Q-value $Q(s, a)$ of the state-action tuple (s, a). If all Q-values are known, selecting the optimal decision is as simple as taking the action with the highest value. However, since the correct Q-values are usually unknown, their estimates are adapted after every timestep, to better account for the obtained reward [38].

For a finite number of discrete states and actions, the Q-values can be represented as a table. On the other hand, for continuous state spaces, we require function approximators, such as Deep Neural Networks (DNNs), to represent the Q-values. This combination of RL techniques and DNNs for function approximation is termed Deep Reinforcement Learning (DRL). For example, in the Deep-Q-Learning (DQL) algorithm, a DNN maps from the continuous state space to the Q-values of a set of actions [10]. For the DQL algorithm, the Q-values of all available actions have to be predicted. This prohibits very large or continuous action spaces. An alternative class of algorithms are the so-called Policy Gradient (PG) methods, which support the utilization

of continuous action spaces. Popular PG algorithms include A3C [39] and DDPG [40].

In the Soft Actor-Critic (SAC) algorithm [41], one DNN—called the critic—is trained to predict the Q-values for a state-action tuple (as in [10]). A second DNN—called the actor—is then used to approximate the Boltzmann distribution over the predicted Q-values of the available actions. At inference time it is therefore sufficient to sample an action from the distribution computed by the actor. In order to use the backpropagation algorithm [42] for the training of the actor, SAC uses a differentiable reparameterization of the sampling operation [43].

III. RELATED WORK

The high relevance of the traffic control problem and the proven ability of RL to solve complex control problems led to a broad range of publications that combine the two topics [44]. The emergence of connected automated vehicles has led to a relatively new research field of traffic control for connected automated vehicles, see, e.g., [45]–[50]. Our study focuses on conventional (non-automated) vehicles.

Approaches differ in the implementation of the traffic MDP—in terms of the utilized state and action space as well as the reward function and the RL learning algorithm. The MDPs differ greatly in terms of the utilized state spaces, which define the knowledge upon which the agent can base its control decisions. Some models assume little information about the current traffic state. For example, [13] uses information from inductive loop sensors in the pavement and could therefore easily be implemented in a contemporary traffic system. Other models assume intricate knowledge of the whereabouts of individual vehicles that could, for example, be inferred over a V2I communication interface. However, none of the existing studies has compared the traffic light system performance for different state spaces. To the best of our knowledge, our present study is the first to quantify in detail the benefits of V2I communication for RL based traffic light control. In particular, our study isolates the performance effects of V2I communication by employing exactly the same RL algorithm for both the holistic agent and the agnostic agent (whereby the two agents differ only in their state spaces).

There is no unique way of defining the action space for controlling a traffic light. However, the framing of the action space can strongly influence the speed of convergence. In [51], for example, the agent can allocate the relative durations of phases while the phase sequence and full length of a cycle are predefined. Such a narrow framing strongly limits the range of behaviors that the system can exhibit, but may speed up convergence, as the search space of the policy is relatively small. In [52] (and in similar variations in [53], [54]), the agent decides in every timestep which phase to show next. This leads to a broader spectrum of behaviors (especially if the length of the interval between timesteps is short) but may converge slowly.

The existing studies also have used distinct RL approaches to solve the respective traffic control MDP. Besides differences in the utilized algorithm, they differ particularly in terms of their use of DNNs as function approximators (DRL) and the application of Multi-Agent Reinforcement Learning (MARL). MARL approaches cope with the high complexity of the simultaneous control of multiple intersections by dividing the task across multiple agents so that, for example, each intersection is controlled by an individual learning agent. In a simple setting, the agents optimize some local reward function of the intersection, e.g., [55]. In more complex settings, higher-order coordination algorithms, such as coordination-graphs and game-theoretic methods, e.g., max-plus, coordinate the actions of multiple agents, e.g., [56]. Building on the seminal MARL studies [13], [52], [55], [57]–[61], recent studies have further substantially advanced MARL computational and decision making strategies, see [62]–[67]. General related computational frameworks for RL control applications have been explored for multi-objective decision modeling in [68] and for a hybrid fuzzy and RL control in [69].

The different utilized reward functions encapsulate the agents' diverse goals. Popular goals are the maximization of velocities, e.g., [51], or the throughput at an intersection, e.g., [13], and the minimization of delays, e.g., [70]–[72]. In MARL approaches, an important distinction is between local and global reward functions. In a local approach, every agent optimizes the performance at its own intersection. In contrast, a global reward function includes metrics from all intersections so that every agent is concerned with the performance across the entire network. This encourages cooperation between agents, especially if good performance of one agent diminishes the performances of nearby agents. For example, an agent using a local reward function could send vehicles to an already overcrowded neighboring intersection, minimizing its delays locally, but deteriorating global performance.

Conceptually, our study is closely related to the recent study [73] on traffic signal control with varying portions of V2I enabled vehicles (ranging from 0 to 100%) at a *single intersection*. (The related study [74] examined the estimation of traffic state with probe vehicles.) Complementary to [73], we examine the performance impact of traffic state information collected via V2I communication for road networks consisting of *multiple intersections*. Essentially, we consider the extreme cases of 0% of V2I enabled vehicles in [73] (roughly equivalent to our agnostic agent) and 100% of V2I enabled vehicles in [73] (roughly equivalent to our holistic agent). We rigorously evaluate these two extreme cases for a variety of multi-intersection road networks ranging from an arterial main road with side roads to a 15-intersection patch of the Barcelona, Spain, road network.

For completeness, we note that V2I control in non-lane based heterogeneous traffic scenarios has been studied in [75]; in contrast, we consider lane-based traffic. Bus holding time control has been examined in [76]; we focus on conventional automobile traffic.

TABLE 1. State spaces of agnostic and holistic agents.

Feature	Agents	
	Agnostic	Holistic
• Current phase of all traffic lights (phase ID & period ID)	✓	✓
• Time passed since the last change	✓	✓
• Traces of all phases	✓	✓
• Positions of vehicles closest to an intersection (road ID, lane ID & distance to intersection)		✓
• Velocities of vehicles closest to an intersection		✓
• Number of vehicles on each lane		✓
• Average velocity of vehicles on each lane		✓

TABLE 2. Parameter values of the traffic environment. Numbers in brackets show differing parameters of the l'Antiga Esquerra de l'Eixample scenario.

Parameter	Value
• Road length	300 m (various)
• Speed limit	20 m/s (various)
• Number of lanes per road and direction	3 (various)
• Minimal green period	5 s
• Maximal green period	100 s
• Number of distinct phase options	8 (various)
• Length of amber period	5 s (4 s)
• Length of all-red period	7 s (2 s)
• Number of individ. observed vehs. per road	10
• Horizon (length of episode)	1 hour/2 hours
• Length of simulation timestep	1 s
• Initial vehicle velocity	5 m/s

IV. DRL FOR TRAFFIC CONTROL

In theory, the ability of a traffic system to obtain detailed information about individual vehicles through a V2I communication link should enable better control decisions and mitigate congestion in the traffic network while upholding high safety standards. In practice however, distilling this large data stream into sensible control decisions is highly complex and traditional traffic control algorithms are unfit to do so. DRL offers a possible solution to the challenges posed by the flood of real-time traffic data. Its proven ability to learn approximate solutions to complex problems from data can enable intelligent traffic control under the knowledge of detailed state information (see Section III). This section introduces the MDP that we developed in this study, including states, actions, and rewards, as well as the DRL algorithm to learn the control of the traffic environment.

A. STATE SPACE

The scope of this study is to evaluate the benefit of providing the traffic infrastructure with detailed traffic state information. The state space of an RL agent defines the knowledge on which it can base its decisions. We therefore compare the performance of two agents with two different state spaces: a state space that is agnostic to the current traffic situation and a state space that is able to observe individual vehicles in the traffic network through V2I communication and thus has a holistic view of the current traffic state. Table 1 summarizes the features included in the state spaces of the two agents.

1) AGNOSTIC AGENT

As the agnostic agent has no means to communicate with vehicles in the traffic network, states are limited to information that is internal to the traffic infrastructure. Most importantly, this includes the signal that is currently shown by the traffic lights. Every intersection has a number of allowed phases which can be identified by a unique phase ID. To further describe the current signal at an intersection, the period ID shows whether the traffic lights currently show the selected phase, the respective amber phase, or the all-red phase. In addition, the agent features the time since the last phase change and a trace for every phase, which increases while the respective phase is activated and slowly decays while it is not. This acts as a memory that lets the agent have some notion of the recent history of activated phases. Note that the traces could be dropped if the agent would employ a model with a memory (such as a Recurrent Neural Network [77]).

We note that our agnostic agent does not directly exploit loop sensors in the road lanes. This is because most currently deployed traffic control algorithms use the loop sensors not to react to individual vehicles, but rather to compute traffic statistics that are then the basis for the traffic signalling. Since we trained the agnostic agent for a prescribed demand, our agnostic agent implicitly uses these traffic statistics.

2) HOLISTIC AGENT

Through a V2I communication interface, the holistic agent receives detailed information about the state of the traffic network. In addition to the features of the agnostic agent, the holistic agent therefore observes further information from the environment. There are many parameters that an intelligent traffic system could potentially leverage towards better control decisions. For example, with information about the fatigue levels of drivers the infrastructure could increase the duration of the amber period or warn nearby drivers. However, for this study, we limit the observation of the traffic system to the positions and velocities of the approaching vehicles. In particular, we observe all approaching vehicles and represent the positions and velocities of a fixed number (10 vehicles, see Table 2) of approaching vehicles per road in individual entries in the state space (the rest is only represented through summary statistics, see below). The position is encoded by the unique ID of the current road, the current lane, and the distance between the vehicle and the next traffic light; the velocity is the absolute speed along the current road. As conventional DNNs need a fixed length input vector, the agent observes the exact location and velocity of a fixed number of vehicles per intersection. If more vehicles are approaching an intersection, the infrastructure observes only the vehicles closest to the intersection. If there are fewer vehicles, the vector is zero-padded. To account for not individually observed vehicles, the state vector also features the

number of approaching vehicles and their average velocity for every road.

B. ACTION SPACE

We let the agent decide in every timestep (*i*) the phase to show, and (*ii*) the display duration. One timestep is defined as one second of simulated time. If the current phase has already been shown longer than a chosen duration (at any decision timestep while a given phase is displayed), then the newly chosen (different) phase will be displayed next. This allows for a broad range of different behaviors while leading to faster convergence than other action spaces we tested. We limit the green phase duration (green period) to be between 5 and 100 seconds and set the amber and all-red periods to fixed values, see Table 2. All available phase options consist of only compatible streams, making the agent's actions inherently safe. Of course, before showing the new phase, the traffic light has to go through an amber phase and an all-red phase to comply with safety regulations. An exemplary set from which an action can be sampled is shown in the lower left of Fig. 3 for a scenario of two controlled traffic lights.

C. REWARD FUNCTION

The reward function defines the goal that the agent strives to achieve. In traffic management, objectives may be manifold: safety, efficiency, environmental sustainability, comfort, and fairness are just some of the possible measures to optimize. Note that some features could be easier to quantify than others and therefore make for a better reward function.

In this study, we mostly use the average velocity of all vehicles as the reward function, since it is a popular measure to quantify the efficacy of a traffic network. Note that the agnostic agent cannot know the average velocity of vehicles in the network as it does not feature a V2I communication interface. The agnostic system therefore would have to be trained in simulations before deployment to the real world, or an external estimator of the average velocity could provide an approximate reward function to learn from a live system. In the simulation experiments in Section V, the agnostic agent is provided with the actual velocities, thus simulating an external estimator. On the other hand, the holistic agent can easily compute the average velocities of vehicles, which are transmitted via a V2I interface, and can thus be trained on a deployed system.

We also experiment with a composite reward function that consists of the average velocity, the average flow rate (the percentage of vehicles that are moving), CO_2 emissions in the traffic network, and stress level of drivers (which, according to [60], is a quadratic function of the time that drivers spent waiting in the recent history). These four factors are weighted equally to form a composite reward function. In one scenario, we study the influence of different reward functions on the resulting policy.

D. LEARNING ALGORITHM

We employ the SAC algorithm with a mixed (discrete and continuous) action space which we found to speed up convergence of the RL agent learning of the traffic control policy. As the original publication [41] intended the SAC algorithm to be used to learn purely continuous control policies, we had to adapt it to be able to cope with our mixed continuous-discrete action space. As suggested in [78], we used the Gumbel-softmax distribution [79], [80] to reparameterize the discrete action choices. To provide better gradients and stabilize learning, we used n-step bootstrapping ($n = 5$) to train the critic and let it predict a distribution of possible Q-values instead of a single Q-value, as proposed in [81]. As the discrete distribution described in [81] requires the predicted Q-values to be bounded, we scale all rewards to be in the range [0, 1]. We experimented with other modifications to the original SAC, e.g., weight decay [82], target policy smoothing [83], and prioritized experience replay [84]; however, these modifications did not improve the learning performance. As the utilized traffic simulator is rather slow, we chose to decouple experience collection and learning. While one process can learn from the experience in a replay buffer, several environments are simulated in parallel to collect new experience and add it to the buffer. Fig. 3 shows the agent-environment interaction loop for a simple traffic network of two connected intersections. Note that we do not display the target networks, that frequently copy the weights of the three DNNs, nor the replay buffer for storing past experience.

We train all DNNs with the Adam optimizer [85]. The two critic networks are learned using a fixed learning rate. In contrast to the two critics, we slowly adapt the learning rate of the actor network so as to approximately constrain the D_2 metric (the sum of the KL-divergence from the old to the new policy and vice versa) of the probabilistic policy before and after every learning step to some predefined value. This mitigates the risk of excessive policy changes which can lead to a sudden significant performance decrease, referred to as policy-breaking. The combination of the second-order gradient approximation of the ADAM optimizer, alongside the after-the-fact adaption of the learning rate to match the desired KL-divergence, could be considered a very coarse approximation of a trust region approach, such as TRPO [86]. The full algorithm is shown in the Appendix.

This study focuses on the traffic engineering aspects rather than the reinforcement learning aspects. Accordingly, we report the results of the trained system (and not the training process of the RL agent). We envision that our agents are trained in simulation, and not in the real world. We trained all agents in simulations until the total undiscounted reward per episode as well as the two loss functions of the Q-function and the policy plateaued.

V. PERFORMANCE COMPARISON

We compare the performance of the holistic agent, which corresponds to the availability of the current traffic state via

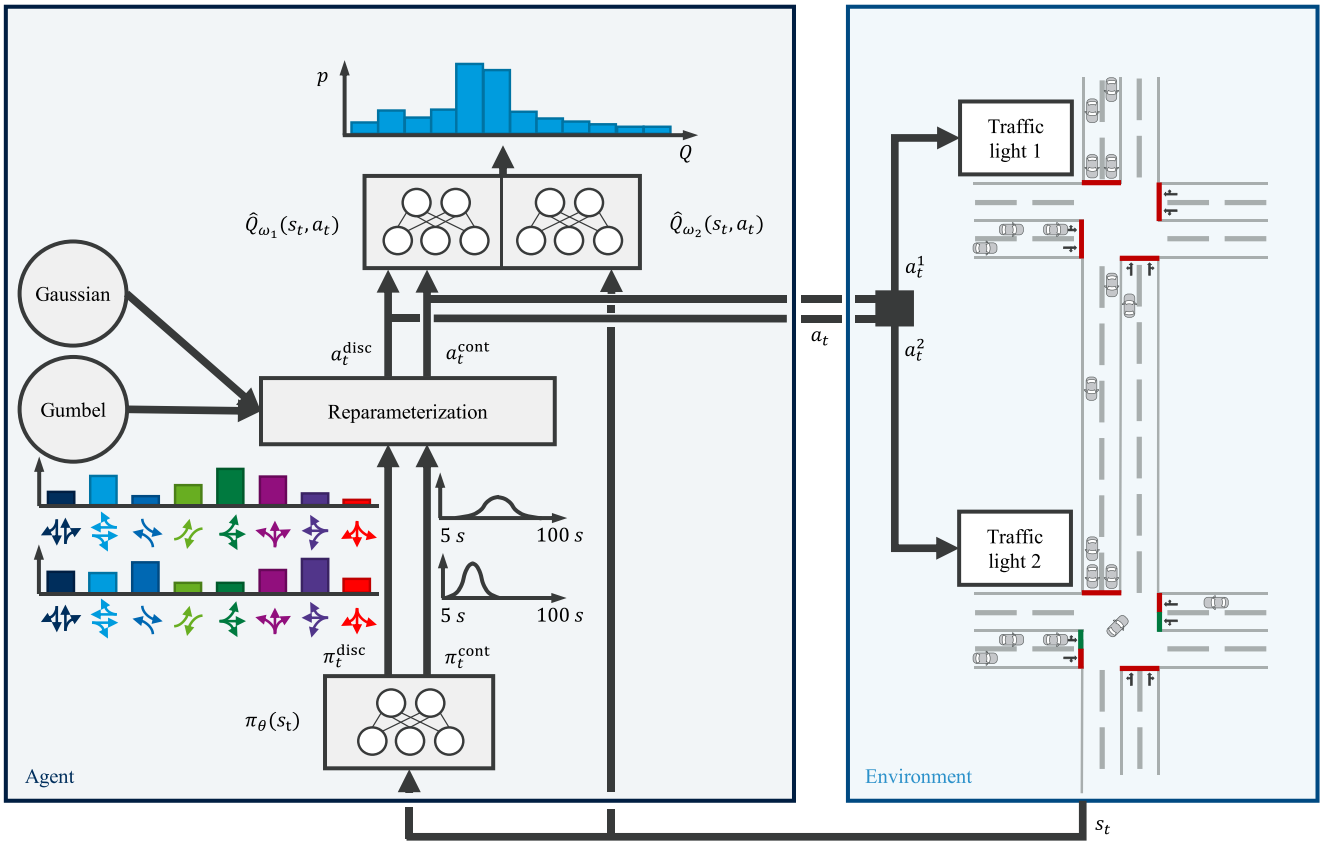


FIGURE 3. Full agent-environment interaction loop for a traffic setting of two connected intersections with eight distinct phase options each. The policy network computes a probability distribution over actions, based on the agent's state. In the reparameterization block, a concrete action is sampled, which is then executed in the environment. The sampled action and the state are used by two Q-functions to predict two probability distributions over action-values. The more pessimistic of the two Q-functions is then used to optimize the policy.

a V2I interface, to the agnostic agent, which corresponds to the lack of the current traffic state, i.e., the absence of a V2I interface. We conduct the performance comparison through simulation experiments in the Urban MObility (SUMO) [87] open-source traffic simulation environment. SUMO is a microscopic simulator, meaning that the dynamics of individual vehicles are explicitly modeled. This is necessary as we want the holistic agent to observe the positions and velocities of individual vehicles. We employed Dijkstra's routing algorithm with edge weights set to the moving average of the travel times for the individual roads, the Intelligent Driver (IDM) car-following [88], and the LC2013 lane-changing model [89]. We employed the default SUMO emission model HBEFA3, simulating a gasoline driven Euro norm 4 passenger car [90], to calculate the CO_2 emissions. To interface the SUMO simulation we use Flow [91]. The Flow framework can be used to model a diverse range of RL problems in the domain of traffic systems, ranging from learned signaling of the traffic infrastructure to the control of individual vehicles. For the RL implementation we use ptan [92] and PyTorch [93] for the realization and training of NNs.

As we want to control traffic lights in this study, we do not make use of the possibility to control vehicles with RL. Instead, all vehicles use SUMO's default controllers for routing, car-following, and lane-changing, as outlined above. The simulation can be entered and left through all roads on the border of the simulated traffic system. Each combination of entry and exit points (except for combinations with identical entry and exit roads) is assigned a Poisson process that generates new vehicles with a predefined spawn rate. The sum of all spawn rates, measured in vehicles per hour (vehs/h), is called the demand or traffic volume. We acknowledge that the assumption of Poisson distributed arrivals of vehicles is a strong one, as vehicles would arrive in waves due to the signaling of adjacent traffic lights in a real traffic network. However, Poisson distributed vehicle arrivals are a common assumption in traffic light control studies, see, e.g., [8], [44], [73].

We do not explicitly model the V2I communication channel. In a more sophisticated approach, we could include a network simulator to account for the dynamics of the information transmission. However, the latency of modern communication technology (on the scale of milliseconds) is small compared to the timescale of traffic light systems with a time-resolution of one second for traffic light phase decision making. A sophisticated communication model would therefore not affect the outcome of the traffic simulations.

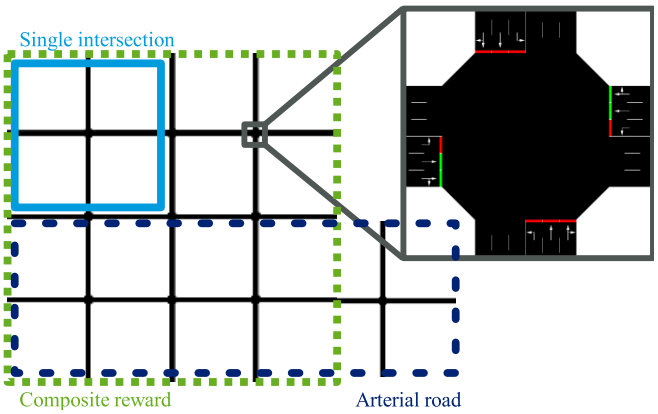


FIGURE 4. Experimental setting of comparisons in Sections V-A–V-D. Colored boxes show the road networks for the single intersection scenario (light blue, Section V-A), the arterial road scenario (dark blue, Sections V-B and V-C), and the composite reward scenario (green, Section V-D).

TABLE 3. Parameter values of the RL algorithm in simulation experiments.

Category	Parameters
Training	• Discount factor: $\gamma = 0.99$ • Batch size: $M = 256$ • Actor learning rate (initial): $\alpha_\pi = 10^{-3}$ • Critic learning rate: $\alpha_Q = 10^{-3}$
Replay buffer	• Buffer size: $R = 10^6$ • Number of initial transitions before learning starts: $B = 10^4$
n-step bootstrapping	• Steps: $N = 5$
Entropy regularisation	• Discrete scaling factor: $\varepsilon_{\text{disc}} = 0.5 \rightarrow 0.01$ (annealed) • Continuous scaling factor: $\varepsilon_{\text{cont}} = 0.01 \rightarrow 0.001$ (annealed)
Distributional Q-value	• Lower bound: $Q_{\min} = 0$ • Upper bound: $Q_{\max} = 100$ • Number of bins: $L = 101$
Target networks	• Update interval: $T_{\text{target}} = 1000$
Reparameterisation	• Gumbel temperature: $G = 2/3$ • Minimal standard deviation of Gaussian: $\sigma_{\min} = 0.005$ • Maximal standard deviation of Gaussian: $\sigma_{\max} = 0.5$
Learning rate adaption	• Target D_2 metric: $D_{\text{target}} = 0.005$ • Parameter of proportional adaption controller: $P = 1$ • Minimal learning rate: $\alpha_{\min} = 10^{-6}$ • Maximal learning rate: $\alpha_{\max} = 10^{-2}$
Distributed experience	• Number of environments simulated in parallel: $K = 3$

In the first four experiments, we consider a generic road infrastructure, as depicted in Fig. 4. We use North-South and East-West roads with three lanes per direction, whereby the leftmost lane allows only left turns, the center lane allows only straights and the rightmost lane allows straights and right turns. At each intersection, the agent can choose from the eight phase options in Fig. 1. In the first experiment (Section V-A), only a single intersection is controlled; whereas in the following two experiments (Sections V-B and V-C), we consider an arterial road of four connected intersections. In the fourth experiment (Section V-D), we study a grid network of three by three intersections (see Fig. 4). In the last experiment (Section V-E), we simulate a patch of the l'Antiga Esquerra de l'Eixample neighborhood in Barcelona, Spain; whereby, the available phase choices correspond to the actual phases of the traffic lights. This data is extracted from OpenStreetMap.org. Fig. 5 shows the simulated road network and the location of the simulated patch on a city map of Barcelona. The traffic environment parameters are summarized in Table 2.

In ML—and particularly DRL—it is often hard to assess the performance of the current policy and to verify that the algorithm has converged to a good solution, as the optimal performance is usually unknown. Longer training times, different hyperparameter choices, or other initial weights of an NN may lead to a better solution. For the results we present here, each learning process was run until the performance of the respective policy did no longer change for a significant amount of time. To mitigate the risk of having found an inferior solution, we ran every learning process several times with a varying set of hyperparameters. Table 3 summarizes the values of the RL system parameters.

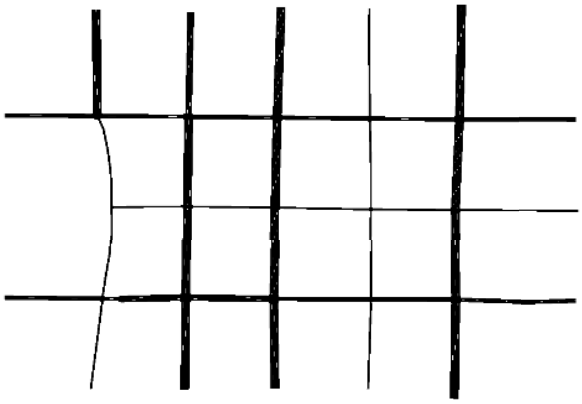
A. SINGLE INTERSECTION

In the first experiment, we tested the atomic setting of a single isolated intersection. We let the spawn rates of the Poisson processes be equal, meaning that average arrival rates for all four approaches are equal and every spawned vehicle is equally likely to take a left turn, go straight, or take a right turn at the intersection. Fig. 6 shows the average velocities of vehicles for the agnostic and the holistic agents for different demands. We also plot the results of an optimized fixed-cycle strategy, which gives the right of way to the different afferent lanes in a round-robin fashion. The optimal phase durations and the best phase cycle (of the two cycles in Fig. 1) were deduced in a brute-force approach, which is feasible for a single isolated intersection, but becomes prohibitive for the larger road networks examined in the subsequent sections.

The single intersection evaluation in Fig. 6 indicates that the agnostic agent can find an equally good solution as the optimal fixed-cycle strategy (within the constraints of the two cycles in Fig. 1). This indicates that the DRL algorithm can find equally good solutions as traditional optimization methods. In subsequent experiments, where finding the optimal fixed-cycle solution is infeasible due to the exponentially increasing number of possible solutions, we only consider the agnostic agent as benchmark for evaluating the holistic agent.

The holistic agent significantly outperforms its agnostic counterpart for low demands. For high demands, however, the advantage of V2I diminishes; for a demand of 3000 vehicles/hour, the holistic agent increases the average vehicle velocity by only 7% compared to the agnostic agent. Analysis of the resulting policies (which is not included due

(a) Map of Barcelona, taken from OpenStreetMaps.



(b) The generated SUMO network.

FIGURE 5. The map in part (a) shows the location of the simulated road network in Barcelona, Spain. Part (b) shows the corresponding 15-intersection road network of the l'Antiga Esquerra de l'Eixample scenario.

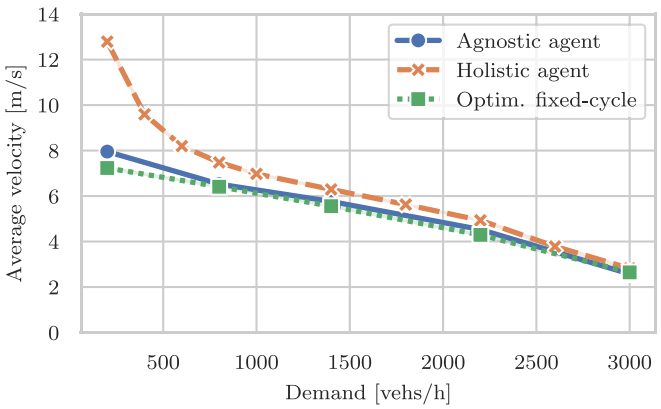


FIGURE 6. Comparison of the average vehicle velocities for the holistic agent, the agnostic agent, and the optimized fixed-cycle strategy for different demands in the single intersection scenario. The holistic agent reliably outperforms the other approaches. The advantage of V2I is especially pronounced for low demands. Shaded areas show the 95% confidence intervals.

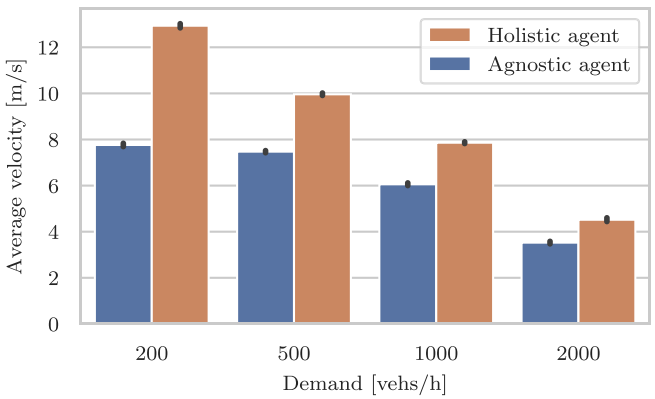


FIGURE 7. Comparison of the average vehicle velocities for the holistic and the agnostic agent for different demands in the arterial scenario. The holistic agent surpasses its agnostic counterpart for all demand settings. Candlewicks show the 95% confidence intervals.

to space constraints) has indicated that the excellent holistic agent performance in low-demand settings is due to the ability of the holistic agent to react to individual vehicles approaching the intersection. Many vehicles can therefore cross the intersection without stopping, as the traffic light system grants the right of way to the respective approaches just in time. For very dense traffic, it is no longer possible to react to individual vehicles, resulting in a decreased impact of the holistic knowledge of the current traffic state.

Note that the agnostic agent needs to learn a separate DNN for every demand setting. The holistic agent, on the other hand, can learn a single DNN for all settings, as it adapts to the current demand. Even though the holistic agent was trained on the same demands as the different agnostic agents (200, 800, 1400, 2200, and 3000 vehs/h), we chose to evaluate the learned model also for other demands (400, 600, 1000, 1800, and 2600 vehs/h) to showcase the holistic agent's ability to generalize to previously unseen demand settings. In a real-world setting, where demands are not steady and hard to accurately quantify, the holistic agent may therefore outperform the other approaches by an even larger margin due to its ability to seamlessly manage a wide range of different demands with a single model.

B. ARTERIAL ROAD

Next, we simulated an arterial road of four connected intersections, as shown in Fig. 4. In contrast to the previous setting, the agent now needs to coordinate the signaling of the traffic lights in order to ensure fluent traffic, making the optimization problem harder. The long arterial road is considered to be a busy main road, whereas the other roads are calm side roads. This is modeled by letting more vehicles enter and leave the simulation on the main road than on the side roads. Fig. 7 shows the average velocities for the two different agents for different demands. The ability of the holistic agent to observe vehicles in the traffic network enables the holistic agent to significantly outperform the agnostic agent. The performance gains of the holistic agent are especially prominent for low demands (67% higher avg. velocity for 200 vehs/h) and decrease for high demands (28% higher avg. velocity for 2000 vehs/h).

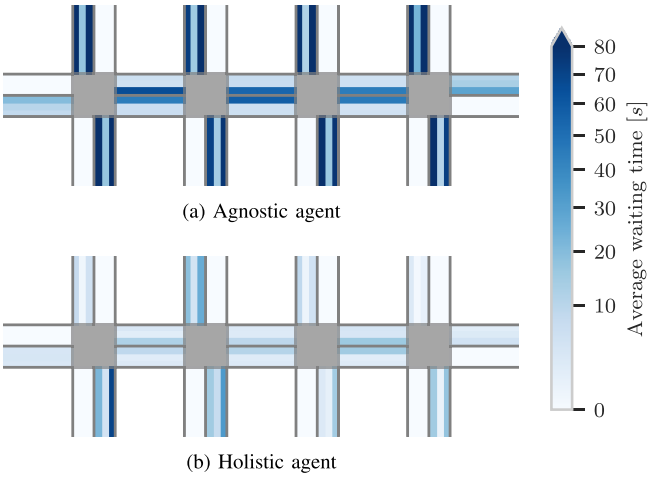


FIGURE 8. Average waiting times for every lane in the arterial scenario for the two different agents and a demand of 200 vehs/h. The holistic agent strongly reduces waiting times of vehicles that enter or leave the main road.

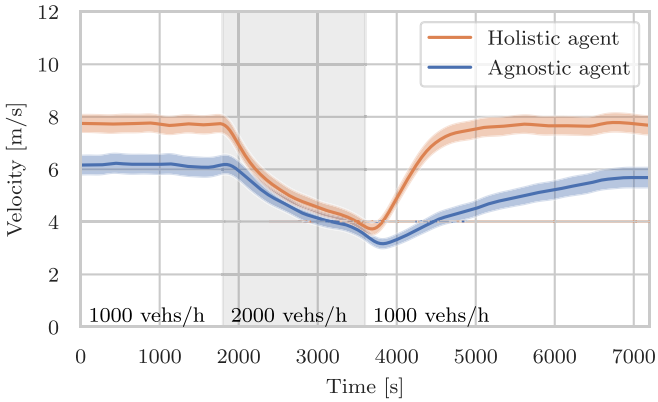


FIGURE 9. Average velocities of vehicles for a sudden inflow in the arterial road scenario, in that the demand is suddenly doubled for 30 minutes and then is reduced back to its former value. The holistic agent reliably outperforms the agnostic agent and recovers significantly faster from the unexpected inflow. Shaded areas show the 95% confidence intervals.

Fig. 8 shows the average time that vehicles spend waiting for every lane of the traffic network for a demand of 200 vehs/h. The knowledge of individual approaching vehicles on side roads allows the holistic agent to grant the right of way to streams entering and leaving the main road when vehicles are approaching and the current traffic situation on the main road allows for a red light. Waiting times are therefore very low on the main road and slightly longer on the side roads. The agnostic agent, on the other hand, grants most green time to the main road, creating "green waves", but rarely gives the right of way to the side roads. This results in relatively fluent traffic on the main road but long waiting times on side roads.

C. SUDDEN INFLOW

In this experiment, we study the effect of unsteady demands. We once more consider the arterial road setup, however instead of investigating the average velocities for different steady demands, we analyze the average velocity (with averaging over 100 independent simulation replications for each 1 second timestep) over time for fluctuating traffic. We simulated the traffic network for two hours, whereby the first 30 minutes have a moderately high demand of 1000 vehs/h, then a very high 2000 vehs/h demand for 30 minutes, and finally 1000 vehs/h for the remaining hour. This setting could, for example, simulate the sharply increased traffic volume after an important sports event, such as a high-profile football match. Fig. 9 shows the average velocity of vehicles over time for the agnostic and the holistic agent.

As before, the holistic agent reliably outperforms its agnostic counterpart. As expected, in the first 30 minutes, the average velocities of both agents correspond to the average velocities of the previous experiment (compare Fig. 7). The sudden increase in traffic volume results in a steady decline of the average velocities, as queues fill up and traffic becomes congested. Interestingly, the velocities fall to a level that is even lower than the measured average velocities for a demand of 2000 vehs/h in the previous experiment. After decreasing the demand back to a moderate level, congestion dissolves and average velocities return to their former levels. However, the ability of the holistic agent to observe its surroundings leads to a significantly faster recovery.

D. COMPOSITE REWARD FUNCTIONS

An essential feature of the described RL methods is that they do not require explicit modeling of the system dynamics. This means that the RL agent can conveniently optimize any numerical reward function, including a weighted combination of several performance metrics. In contrast, most traditional traffic optimization methods are explicitly designed to optimize a single metric. In this experiment, we study the effects of different reward functions. We compare the average velocity, the flow rate (percentage of vehicles that are moving in the road network), the CO_2 emissions (total emissions in the road network), and the stress level of drivers of a system that—as in previous experiments–optimizes only the average velocity, against a system that optimizes a weighted combination of the four metrics. Both systems utilize a V2I communication channel. Fig. 10 shows the distributions of the metrics as well as their pairwise correlations obtained from 100 independent simulation replications, each simulating the road network for one hour with the same moderate traffic demand of 1000 vehicles per hour.

The composite reward function results in an agent that performs on par or outperforms the single-metric agent for all considered metrics. Through shorter green times, the number and duration of full stops in the traffic network is reduced, resulting in approximately 3% higher flow rate, lower stress levels (reduced from a mean of 16.9% of the average stress level down to a mean of 13.9%), and 3% lower CO_2 emissions. The average velocities for both reward functions differ only marginally (approx. 1%). Also, all the variances are reduced by the composite reward function.

RL clearly shows the potential to jointly optimize the manifold objectives of modern traffic systems. However, due to the strong correlations of the four performance metrics in this

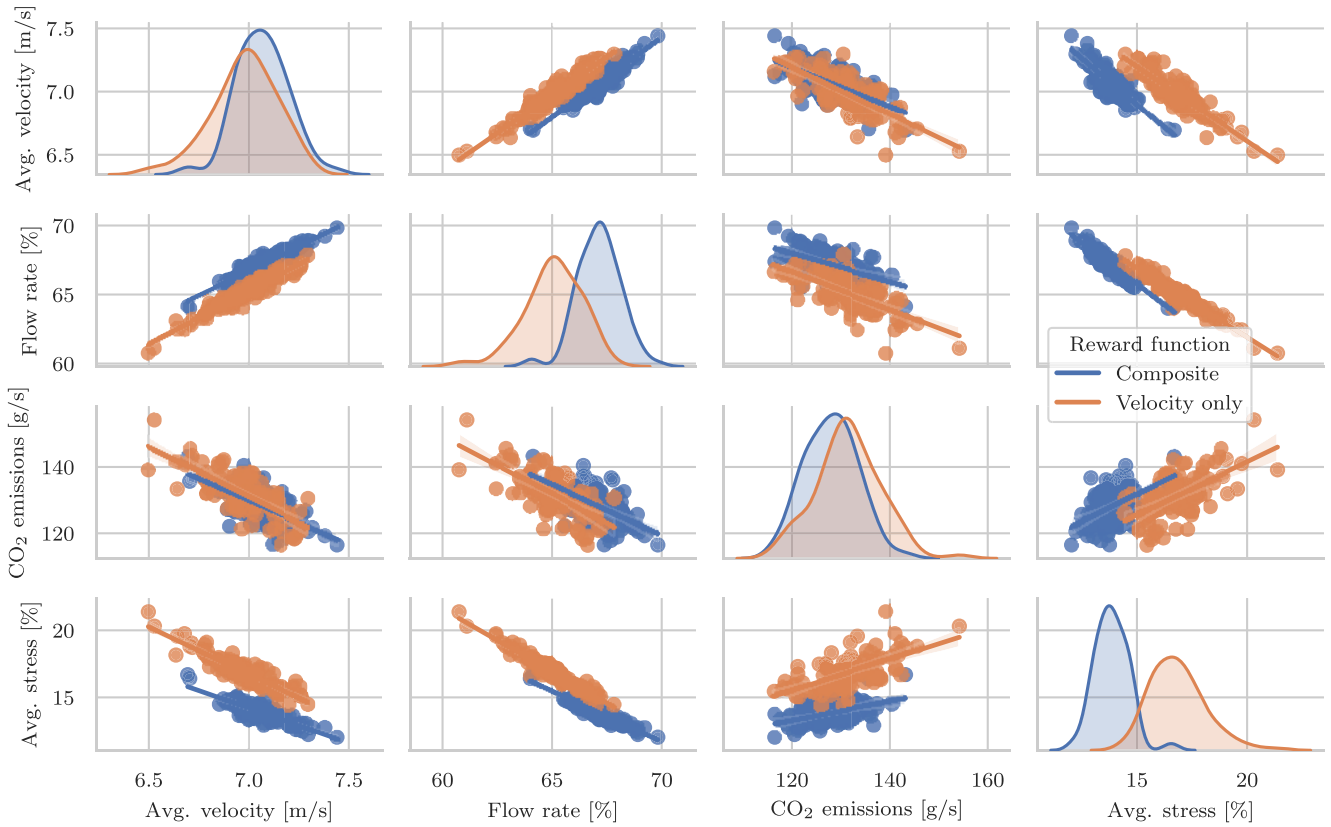


FIGURE 10. Comparison of different performance metrics for the composite reward function scenario, where two holistic agents are trained with different reward functions. One reward function is based solely on the average velocity of the vehicles; the other is a weighted combination of the four metrics. The plot shows distributions of the four metrics as well as their pairwise correlations. The RL approach successfully optimizes the composite reward function. However, the strong correlations of the four metrics results in very similar outcomes for both rewards.

experiment, the resulting policies of the two reward functions differ only slightly. Future research should therefore investigate the joint optimization of conflicting performance indicators, such as velocity and pedestrian safety, forcing the agent to trade off the different objectives. For a real-world implementation, the employed reward function should be selected with the utmost care as an unsuitable quantification of objectives may produce unexpected and possibly dangerous results.

E. CENTRAL URBAN NEIGHBORHOOD EXAMPLE

In the last experiment, we replace the generic traffic network of Fig. 4 with a patch of the l'Antiga Esquerra de l'Eixample neighborhood in Barcelona, as shown in Fig. 5. This central neighborhood consists mostly of residential buildings and offices—with the exception of the large 'Hospital Clinic' on the west end of the road network, giving rise to heavy commuter traffic, especially during the rush hours in the morning and late afternoon. Spawn rates of streets that enter the simulation are proportional to their respective numbers of lanes. Likewise, the probability of a street being the destination of a vehicle is proportional to its number of lanes. As the utilization of the composite reward function led to little difference in the resulting policy, we again compare the performance of the agnostic and the holistic agents that optimize only the average velocity of vehicles (as in experiments A-C). However, we here compare more performance metrics than only the average velocity. Fig. 11 shows the average velocity of vehicles, the average flow rate, average CO_2 emissions in the system, the average stress level of drivers, the average time that vehicles need to traverse the traffic network, and the average time that vehicles spent waiting at traffic lights during their trip.

The holistic agent outperforms the agnostic agent in terms of all metrics and for all demands. The higher amount of coordination between the traffic lights, as performed by the holistic agent, results in a pronounced advantage for all three demand scenarios. For example, for all three demand scenarios, the holistic agent manages to allocate green time more efficiently and to reduce waiting times by approximately 50%. The lower waiting times result in a higher average velocity and flow rate as well as reduced stress levels. Comparing the reductions of the average trip times (25.36 s for 1000 vehs/h, 26.59 s for 2000 vehs/h, and 27.33 s for 3000 vehs/h) and of the average waiting times (11.99 s for 1000 vehs/h, 15.53 s for 2000 vehs/h, and 17.49 s for 3000 vehs/h), we observe that they are not equal. This difference in reduction of trip times and waiting times shows that the holistic agent not only decreases the average waiting time per stop but also the total number of stops, reducing the need for deceleration and acceleration among vehicles and resulting in more fluent traffic. The decreased acceleration

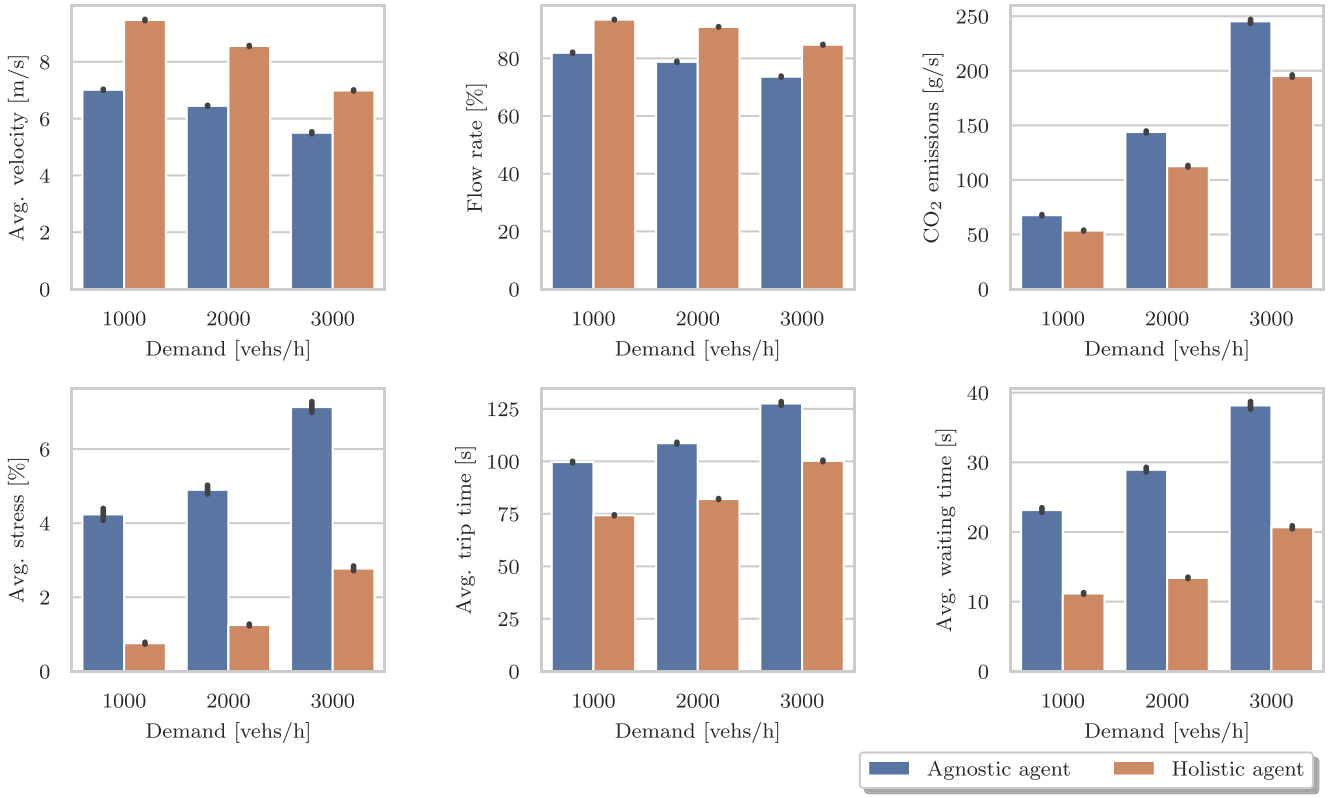


FIGURE 11. Comparison of different performance metrics in the L'Antigua Esquerra de l'Eixample setting for the two agents. The holistic agent consistently outperforms the agnostic agent in terms of all six metrics and for all tested demands. Candlewicks show the 95% confidence intervals.

among vehicles alongside the diminished time that vehicles spend on the road result in a roughly 20% reduction of CO_2 emissions, mitigating the environmental repercussions of congestion.

VI. DISCUSSION

We have developed a DRL system that controls one or several traffic lights in a simulated traffic environment. This DRL approach showed to be able to effectively learn the intelligent control of traffic light signaling at multiple intersections from interaction with its environment. We compared the performance of an agnostic agent, that cannot communicate with vehicles in the traffic network, with the performance of a holistic agent, that features a V2I communication interface and therefore knows the positions and velocities of all vehicles.

Our results show the enormous potential of V2I technologies in the mitigation of congestion. Across an extensive range of multi-intersection road network simulations, we showed that a holistic view of the state of the traffic network empowers the traffic system to make highly informed control decisions. This manifests, for example, in higher average velocities, shorter waiting times, lower CO_2 emissions, and reduced stress levels in drivers. By training the holistic agent in scenarios of varying traffic volumes, the holistic agent learns to seamlessly integrate the broad range of behaviors that are needed to navigate the different requirements. The advantage of communication with nearby vehicles showed to be especially pronounced for low traffic volumes, as the agent learns to react to individual vehicles, often allowing cars to traverse the traffic network without ever stopping. The advantage slowly decreases for high traffic volumes since too many vehicles approach the intersections to react to individual vehicles.

The use of composite reward functions enables the joint optimization of multiple performance metrics. Through a V2I interface, individual vehicles may transmit a plethora of different metrics, enabling the design of reward functions that accurately encapsulate the diverse objectives of traffic systems. However, the traffic infrastructure cannot measure most performance metrics directly but has to rely on collaborative information from individual vehicles to construct meaningful reward functions. A further advantage of traffic systems featuring V2I communication is thus the transmission of these performance indicators, such as velocity, waiting time, or CO_2 emissions, from the individual vehicles to the infrastructure.

VII. OUTLOOK

The weak performance guarantees of DRL methods are the main reason for their rare implementation in safety-relevant systems. As traffic light control systems can utilize an action space that is safe by design, e.g., by using only compatible streams for the available phases and enforcing appropriate amber periods, they could be a suitable domain for first pilot projects of DRL for safety-relevant applications. Such a system could be trained in simulation until a reasonable solution is found, and then be deployed into the real world,

Algorithm 1: RL Algorithm for Learning Traffic Signal Control

Input: Batch size M; discount factor γ; replay buffer size R; initial environment steps B; actor learning rate α_π; critic learning rate α_Q; n-step bootstrapping steps N; discrete entropy scaling factor $\varepsilon_{\text{disc}}$; continuous entropy scaling factor $\varepsilon_{\text{cont}}$; target network update interval T_{target}; parameters of the categorical action-value distribution: lower bound $Q_{\min}$, upper bound $Q_{\max}$ and number of discrete bins L.

Initialize network weights $(\theta, \omega_1, \omega_2)$ using Kaiming initialization [95].
Initialize target network weights $(\theta', \omega_1', \omega_2') \leftarrow (\theta, \omega_1, \omega_2)$.
Initialize replay buffer $\mathcal{B}$.
Launch K agents and environments.
while *True* **do**
  if $len(\mathcal{B}) \geq B$ **then**
    Sample minibatch of M transitions of length N from the replay buffer.
    Sample for each transition $a'_{i+N} \sim \pi_{\theta'}(s_{i+N})$.
    Compute and store for each transition $\pi_{\text{old},i} = \pi_\theta(s_i)$
    Compute $\omega'_{\min} = \arg\min_{\omega'_1,\omega'_2} \left(\text{avg}(\hat{Q}_{\omega'_1}(s_{i+N}, a'_{i+N})), \text{avg}(\hat{Q}_{\omega'_2}(s_{i+N}, a'_{i+N}))\right)$.
    Construct the target distributions $Y_i = \left(\sum_{n=0}^{N-1} \gamma^n r_{i+n}\right) + \gamma^N \hat{Q}_{\omega'_{\min}}(s_{i+N}, a'_{i+N})$
      (The target distribution Y_i is constructed by moving the probability mass of each of the discrete bins of the categorical distribution according to the Bellman equation (see [81])).
    Compute actor and critic updates ($\mathcal{H}(\pi)$ is the entropy of the distribution π):

$$\theta \leftarrow \theta - \alpha_\pi \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta \left(- \hat{Q}_{\omega_1}(s_i, a'_i \sim \pi_\theta(s_i)) - \varepsilon_{\text{disc}} \mathcal{H}(\pi_\theta(s_i)) - \varepsilon_{\text{cont}} \mathcal{H}(\pi_\theta(s_i)) \right)$$

$$\omega_1 \leftarrow \omega_1 - \alpha_Q \frac{1}{M} \sum_{i=1}^{M} \nabla_{\omega_1} D_{KL}\left(Y_i \;||\; \hat{Q}_{\omega_1}(s_i, a_i)\right)$$

$$\omega_2 \leftarrow \omega_2 - \alpha_Q \frac{1}{M} \sum_{i=1}^{M} \nabla_{\omega_2} D_{KL}\left(Y_i \;||\; \hat{Q}_{\omega_2}(s_i, a_i)\right)$$

    if $t \;\%\; T_{\text{target}} = 0$ **then**
      Copy parameters to target networks $(\theta', \omega'_1, \omega'_2) \leftarrow (\theta, \omega_1, \omega_2)$.
    end

    Adapt policy learning rate $\alpha_\pi \leftarrow \text{clip}\left[\alpha_\pi - P \cdot \left(\frac{1}{M}\sum_{i=1}^{M} D_2(\pi_\theta(s_i)||\pi_{\text{old},i}) - D_{\text{target}}\right), \alpha_{\min}, \alpha_{\max}\right]$
  end
end

Agent
while *True* **do**
  Observe state s and sample action from policy $a \sim \pi_\theta(s)$
  Execute action a and observe reward r and new state s'.
  Store tuple (s, a, r, s') in replay buffer and delete old entries from the buffer if $len(\mathcal{B}) > R$.
end

where it can keep on learning to further adapt to the actual requirements of the system. For this continued learning, the availability of a V2I interface is imperative as the reward metrics from individual vehicles need to be communicated to the infrastructure.

In this study, we optimized the average velocity of vehicles in the network or a simple combination of a few performance metrics. Such simple reward functions could turn out to be short-sighted when using a more sophisticated simulation or deploying the agent to the real world. We may, for example, integrate additional goals, such as pedestrian safety [67], [94], fairness among drivers, or noise levels in residential areas. Further research needs to identify tangible goals of traffic systems and appropriately quantify and balance them.

Finally, another exciting research direction would be the integration of the bilateral exchange of information between the traffic infrastructure and individual vehicles. For example, giving drivers suggestions on appropriate speed or routes could enable the infrastructure to better handle and distribute traffic and, therefore, use the given road infrastructure more efficiently, further mitigating congestion.

APPENDIX

This appendix presents the RL algorithm for learning of the traffic signal management in Algorithm 1.

REFERENCES

[1] *European Urban Mobility—Policy Context*, Eur. Commission, Brussels, Belgium, 2017. Accessed: Apr. 1, 2020. [Online]. Available: https://ec.europa.eu/transport/sites/transport/files/2017-sustainable-urban-mobility-policy-context.pdf

[2] *Inrix Global Traffic Scoreboard*, Inrix, Kirkland, WA, USA, 2019. Accessed: Apr. 1, 2020. [Online]. Available: http://inrix.com/scorecard/

[3] G. S. Callendar, "The artificial production of carbon dioxide and its influence on temperature," *Quart. J. Roy. Meteorol. Soc.*, vol. 64, no. 275, pp. 223–240, Apr. 1938.

[4] W. Xie *et al.*, "Decreases in global beer supply due to extreme drought and heat," *Nat. Plants*, vol. 4, no. 11, pp. 964–973, Nov. 2018.

[5] L. Chen and C. Englund, "Cooperative intersection management: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 570–586, Feb. 2016.

[6] Y. Jeong, S. Kim, and K. Yi, "Surround vehicle motion prediction using LSTM-RNN for motion planning of autonomous vehicles at multi-lane turn intersections," *IEEE Open J. Intell. Transp. Syst.*, vol. 1, pp. 2–14, 2020.

[7] A. I. Morales Medina, F. Creemers, E. Lefeber, and N. van de Wouw, "Optimal access management for cooperative intersection control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2114–2127, May 2020.

[8] O. K. Tonguz and R. Zhang, "Harnessing vehicular broadcast communications: DSRC-actuated traffic control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 509–520, Feb. 2020.

[9] N. Wu, D. Li, Y. Xi, and B. de Schutter, "Distributed event-triggered model predictive control for urban traffic lights," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 23, 2020, doi: 10.1109/TITS.2020.2981381.

[10] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[11] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[12] I. Popov *et al.*, "Data-efficient deep reinforcement learning for dexterous manipulation," Apr. 2017. [Online]. Available: http://arxiv.org/abs/1704.03073.

[13] S. Richter, D. Aberdeen, and J. Yu, "Natural actor-critic for road traffic optimisation," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2006, pp. 1169–1176.

[14] M. Papageorgiou, "Overview of road traffic control strategies," *IFAC Proc. Vol.*, vol. 37, no. 19, pp. 29–40, Oct. 2004.

[15] R. E. Allsop, "SIGSET: A computer program for calculating traffic signal settings," *Traffic Eng. Control*, vol. 2, no. 13, pp. 58–60, Jun. 1971.

[16] R. E. Allsop, "SIGCAP: A computer programm for assesing the traffic capacity of signal-controlled road junctions," *Traffic Eng. Control*, vol. 17, no. 819, pp. 338–341, Aug. 1976.

[17] R. A. Vincent and C. P. Young, "Self-optimizing traffic signal control using microprocessors: The TRRL 'MOVA' strategy for isolated intersections," in *Proc. 2nd Int. Conf. Road Traffic Control*, Dec. 1986, pp. 102–105.

[18] J. D. C. Little, "The synchronization of traffic signals by mixed-integer linear programming," *Oper. Res.*, vol. 14, no. 4, pp. 568–594, Aug. 1966.

[19] P. B. Hunt and D. I. Robertson, "The SCOOT on-line traffic signal optimisation technique," *Traffic Eng. Control*, vol. 23, no. 4, pp. 190–192, Jun. 1982.

[20] C. Kergaye, A. Stevanovic, and P. Martin, "An evaluation of SCOOT and SCATS through microsimulation," in *Proc. 10th Int. Conf. Appl. Adv. Technol. Transp.*, Jan. 2009, pp. 1166–1180.

[21] J. Henry, J. Farges, and J. Tuffal, "The Prodyn real time traffic algorithm," *IFAC Proc. Vol.*, vol. 16, no. 4, pp. 305–310, Apr. 1983.

[22] P. Chakroborty and A. Das, "Design of traffic facilities," in *Principles of Transporation Engineering*, R. Babuška and F. C. A. Groen, Eds. New Delhi, India: PHI Learn., 2017, ch. 5, pp. 179–204.

[23] F. Arena and G. Pau, "An overview of vehicular communications," *Future Internet*, vol. 11, no. 2, pp. 27–38, Jan. 2019.

[24] H. Joo, S. H. Ahmed, and Y. Lim, "Traffic signal control for smart cities using reinforcement learning," *Comput. Commun.*, vol. 154, pp. 324–330, Mar. 2020.

[25] J. Kim, S. Jung, K. Kim, and S. Lee, "The real-time traffic signal control system for the minimum emission using reinforcement learning in V2X environment," *Chem. Eng. Trans.*, vol. 72, pp. 91–96, Jan. 2019.

[26] W. Liu, G. Qin, Y. He, and F. Jiang, "Distributed cooperative reinforcement learning-based traffic signal control that integrates V2X networks' dynamic clustering," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 8667–8681, Oct. 2017.

[27] C. Han, M. Dianati, R. Tafazolli, R. Kernchen, and X. Shen, "Analytical study of the IEEE 802.11p MAC sublayer in vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 873–886, Jun. 2012.

[28] L. Shi and K. W. Sung, "Spectrum requirement for vehicle-to-vehicle communication for traffic safety," in *Proc. 79th IEEE Veh. Techol. Conf.*, Seoul, South Korea, May 2014, pp. 1–5.

[29] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, "LTE for vehicular networking: A survey," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 148–157, May 2013.

[30] R. Molina-Masegosa and J. Gozalvez, "LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications," *IEEE Veh. Technol. Mag.*, vol. 12, no. 4, pp. 30–39, Dec. 2017.

[31] W. Kellerer, P. Kalmbach, A. Blenk, A. Basta, M. Reisslein, and S. Schmid, "Adaptable and data-driven softwarized networks: Review, opportunities, and challenges," *Proc. IEEE*, vol. 107, no. 4, pp. 711–731, Apr. 2019.

[32] A. Nasrallah *et al.*, "Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 88–145, 1st Quart., 2019.

[33] B. P. Gopal and P. G. Kuppusamy, "A comparative study on 4G and 5G technology for wireless applications," *IOSR J. Elect. Commun. Eng.*, vol. 10, no. 6, pp. 67–72, Nov. 2015.

[34] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, May 2019.

[35] J. A. Cabrera, R. Schmoll, G. T. Nguyen, S. Pandi, and F. H. P. Fitzek, "Softwarization and network coding in the mobile edge cloud for the tactile Internet," *Proc. IEEE*, vol. 107, no. 2, pp. 350–363, Feb. 2019.

[36] Z. Ning *et al.*, "Joint computing and caching in 5G-envisioned Internet of Vehicles: A deep reinforcement learning-based traffic control system," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 5, 2020, doi: 10.1109/TITS.2020.2970276.

[37] Z. Xiang, F. Gabriel, E. Urbano, G. T. Nguyen, M. Reisslein, and F. H. Fitzek, "Reducing latency in virtual machines: Enabling tactile Internet for human-machine co-working," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1098–1116, May 2019.

[38] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, May 1992.

[39] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, Jun. 2016, pp. 1928–1937.

[40] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent.*, Jan. 2016. [Online]. Available: https://arxiv.org/abs/1509.02971

[41] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, Jul. 2018, pp. 1861–1870.

[42] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[43] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent.*, Apr. 2014, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1312.6114

[44] A. Haydari and Y. Yilmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," 2020. [Online]. Available: arXiv:2005.00935.

[45] K. Higashiyama, K. Kimura, H. Babakarkhail, and K. Sato, "Safety and efficiency of intersections with mix of connected and non-connected vehicles," *IEEE Open J. Intell. Transp. Syst.*, vol. 1, pp. 29–34, 2020.

[46] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 7, 2020, doi: 10.1109/TITS.2019.2962338.

[47] Y. Shao and Z. Sun, "Eco-approach with traffic prediction and experimental validation for connected and autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 13, 2020, doi: 10.1109/TITS.2020.2972198.

[48] M. Tajalli, M. Mehrabipour, and A. Hajbabaie, "Network-level coordinated speed optimization and traffic light control for connected and automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, early access, Jun. 5, 2020, doi: 10.1109/TITS.2020.2994468.

[49] H. Yang, F. Almutairi, and H. Rakha, "Eco-driving at signalized intersections: A multiple signal optimization approach," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 9, 2020, doi: 10.1109/TITS.2020.2978184.

[50] M. Zhou, Y. Yu, and X. Qu, "Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 433–443, Jan. 2020.

[51] N. Casas, "Deep deterministic policy gradient for urban traffic light control," Mar. 2017. [Online]. Available: http://arxiv.org/abs/1703.09035.

[52] E. Van der Pol and F. A. Oliehoek, "Coordinated deep reinforcement learners for traffic light control," in *Proc. NIPS Workshop Learn. Inference Control Multi-Agent Syst.*, Dec. 2016, pp. 1–8. [Online]. Available: https://www.fransoliehoek.net/docs/VanDerPol16LICMAS.pdf

[53] W. Genders and S. N. Razavi, "Using a deep reinforcement learning agent for traffic signal control," Nov. 2016. [Online]. Available: http://arxiv.org/abs/1611.01142.

[54] L. Li, Y. Lv, and F. Wang, "Traffic signal timing via deep reinforcement learning," *IEEE/CAA J. Automatica Sinica*, vol. 3, no. 3, pp. 247–254, Jul. 2016.

[55] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, "Reinforcement learning-based multi-agent system for network traffic signal control," *IET Intell. Transp. Syst.*, vol. 4, no. 2, pp. 128–135, Jun. 2010.

[56] B. Bakker, S. Whiteson, L. Kester, and F. C. A. Groen, "Traffic light control by multiagent reinforcement learning systems," in *Interactive Collaborative Information Systems*, R. Babuška and F. C. A. Groen, Eds. Heidelberg, Germany: Springer, 2010, ch. 18, pp. 475–510.

[57] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): Methodology and large-scale application on downtown Toronto," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1140–1150, Sep. 2013.

[58] K. J. Prabuchandran, A. N. Hemanth Kumar, and S. Bhatnagar, "Decentralized learning for traffic signal control," in *Proc. 7th Int. Conf. Commun. Syst. Netw.*, Bangalore, India, Jan. 2015, pp. 1–6.

[59] P. Mannion, J. Duggan, and E. Howley, "An experimental review of reinforcement learning algorithms for adaptive traffic signal control," in *Autonomic Road Transport Support Systems*, vol. 1, T. L. McCluskey *et al.*, Eds. Cham, Switzerland: Springer, 2016, ch. 4, pp. 47–66.

[60] M. Liu, J. Deng, X. Ming, X. Zhang, and W. Wang, "Cooperative deep reinforcement learning for traffic signal control," in *Proc. 6th Int. Workshop Urban Comput.*, Aug. 2017, pp. 1–8. [Online]. Available: https://urbcomp.ist.psu.edu/2017/papers/Cooperative.pdf

[61] D. Garg, M. Chli, and G. Vogiatzis, "Deep reinforcement learning for autonomous traffic light control," in *Proc. 3rd IEEE Int. Conf. Intell. Transp. Eng.*, Singapore, Sep. 2018, pp. 214–218.

[62] T. Chu, J. Wang, L. Codeca, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, Mar. 2020.

[63] A. Hussain, T. Wang, and C. Jiahua, "Optimizing traffic lights with multi-agent deep reinforcement learning and V2X communication," 2020. [Online]. Available: arXiv:2002.09853.

[64] K. Menda *et al.*, "Deep reinforcement learning for event-driven multi-agent decision processes," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1259–1268, Apr. 2019.

[65] F. Rasheed, K.-L. A. Yau, and Y.-C. Low, "Deep reinforcement learning for traffic signal control under disturbances: A case study on Sunway city, Malaysia," *Future Gener. Comput. Syst.*, vol. 109, pp. 431–445, Aug. 2020.

[66] N. Wu, D. Li, and Y. Xi, "Distributed weighted balanced control of traffic signals for urban traffic congestion," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3710–3720, Oct. 2019.

[67] T. Wu *et al.*, "Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8243–8256, Aug. 2020.

[68] J. Jin and X. Ma, "A multi-objective agent-based control approach with application in intelligent traffic signal system," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3900–3912, Oct. 2019.

[69] N. Kumar, S. S. Rahman, and N. Dhakad, "Fuzzy inference enabled deep reinforcement learning-based traffic light control for intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, early access, Apr. 20, 2020, doi: 10.1109/TITS.2020.2984033.

[70] X. Liang, X. Du, G. Wang, and Z. Han, "A deep reinforcement learning network for traffic light cycle control," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1243–1253, Feb. 2019.

[71] S. S. Mousavi, M. Schukat, and E. Howley, "Traffic light control using deep policy-gradient and value-function-based reinforcement learning," *IET Intell. Trans. Syst.*, vol. 11, no. 7, pp. 417–423, Sep. 2017.

[72] L. Shoufeng, L. Ximin, and D. Shiqiang, "Q-learning for adaptive traffic signal control based on delay minimization strategy," in *Proc. IEEE Int. Conf. Netw. Sens. Control*, Sanya, China, Dec. 2008, pp. 687–691.

[73] R. Zhang, A. Ishikawa, W. Wang, B. Striner, and O. K. Tonguz, "Using reinforcement learning with partial vehicle detection for intelligent traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 30, 2020, doi: 10.1109/TITS.2019.2958859.

[74] C. N. Van Phu and N. Farhi, "Estimation of urban traffic state with probe vehicles," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 25, 2020, doi: 10.1109/TITS.2020.2975120.

[75] S. D. Kumaravel and R. Ayyagari, "A decentralized signal control for non-lane-based heterogeneous traffic under V2I communication," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1741–1750, Apr. 2020.

[76] G. Laskaris, M. Seredynski, and F. Viti, "Enhancing bus holding control using cooperative ITS," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1767–1778, Apr. 2020.

[77] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[78] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Feb. 2018, pp. 1495–1502.

[79] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. NIPS Workshop Bayesian Deep Learn.*, Dec. 2016, pp. 1–12. [Online]. Available: https://arxiv.org/pdf/1611.01144v2.pdf

[80] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *Proc. NIPS Workshop Bayesian Deep Learn.*, Dec. 2016, pp. 1–20. [Online]. Available: http://arxiv.org/abs/1611.00712

[81] G. Barth-Maron *et al.*, "Distributed distributional deterministic policy gradients," Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.08617.

[82] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. 4th Int. Conf. Neural Inform. Process. Syst.*, Dec. 1991, pp. 950–957.

[83] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, Jul. 2018, pp. 1582–1591. [Online]. Available: http://proceedings.mlr.press/v80/fujimoto18a.html

[84] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. 4th Int. Conf. Learn. Represent.*, Jan. 2016, pp. 1–21.

[85] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, May 2015, pp. 1–15. [Online]. Available: https://arxiv.org/abs/1412.6980v9

[86] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 1889–1897.

[87] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "SUMO—Simulation of Urban MObility: An overview," in *Proc. SIMUL 3rd Int. Conf. Adv. Syst. Simulat.*, Jul. 2011, pp. 63–68.

[88] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E*, vol. 62, no. 2, pp. 1805–1824, Aug. 2000.

[89] J. Erdmann, "SUMO's lane-changing model," in *Modeling Mobility With Open Data*, M. Behrisch and M. Weber, Eds. Cham, Switzerland: Springer, Mar. 2015, pp. 105–123.

[90] S. Hausberger, M. Rexeis, M. Zallinger, and R. Luz, "Emission factors from the model PHEM and HBEFA version 3," Dept. Internal Combustion Engines Thermodyn., TU Graz, Graz, Austria, Rep. I-20/2009 Haus-Em 33/08/679, Dec. 2009.

[91] C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: A modular learning framework for autonomy in traffic," Oct. 2017. [Online]. Available: http://arxiv.org/abs/1710.05465.

[92] M. Lapan, *Deep Reinforcement Learning Hands-On: Apply Modern RL Methods, With Deep Q-Networks, Value Iteration, Policy Gradients, TRPO, AlphaGo Zero and More*. Birmingham, U.K.: Packt Publ., 2018.

[93] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop Autodiff*, Dec. 2017, pp. 1–4.
[94] Y. Zhang, Y. Zhang, and R. Su, "Pedestrian-safety-aware traffic light control strategy for urban traffic congestion alleviation," *IEEE Trans. Intell. Transp. Syst.*, early access, Dec. 4, 2019, doi: 10.1109/TITS.2019.2955752.
[95] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 1026–1034.

JOHANNES V. S. BUSCH received the Diploma degree (Dipl.Ing.) in electrical engineering from Technical University of Dresden, Germany, in 2019, where he is currently pursuing the Ph.D. degree with the Deutsche Telekom Chair of Communication Networks. His research focuses mainly on machine learning algorithms—in particular deep reinforcement learning and deep generative modeling—and their application to complex real-world problems, such as efficient traffic control and the optimization of communication networks.

VINCENT LATZKO (Graduate Student Member, IEEE) received the Diploma degree in electrical engineering from Technische Universität Darmstadt, Germany, in 2017. He is currently pursuing the Ph.D. degree with the Deutsche Telekom Chair of Communication Networks, Technical University of Dresden, Germany. His current research interests include efficient compression, softwarized network infrastructure, and reliable low latency communication. His main research focus is on joining artificial intelligence in multi-access edge computing and optimized vehicular traffic flows.

MARTIN REISSLEIN (Fellow, IEEE) received the Ph.D. degree in systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 1998. He is currently a Professor with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, USA, and also an External Associate Investigator with the Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden, Germany. He is currently an Associate Editor-in-Chief of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS and a Co-Editor-in-Chief of the *Optical Switching and Networking*.

FRANK H. P. FITZEK (Senior Member, IEEE) received the Dipl.Ing. degree in electrical engineering from the University of Technology and Rheinisch-Westfälische Technische Hochschule (RWTH), Aachen, Germany, in 1997, the Ph.D. (Dr.Ing.) degree in electrical engineering from Technical University, Berlin, Germany, in 2002, and the Doctor Honoris Causa degree from the Budapest University of Technology and Economy in 2015. He is currently a Professor and the Head of the Deutsche Telekom Chair of Communication Networks with Technical University of Dresden, Germany, coordinating the 5G Lab Germany. He is the Spokesman of the DFG Cluster of Excellence Centre for Tactile Internet with Human-in-the-Loop (CeTI).