ABSTRACT

The transportation industry is one of the main contributors to global warming since it is responsible for a quarter of greenhouse gas emissions. Due to society’s crucial dependence on fossil fuels and the rapid increase in mobility demands, the reduction of global vehicle emissions evolved into a significant challenge. In the urban transportation areas, signalized intersections can be considered the main bottlenecks in mitigating congestion and, therefore, vehicle emission. Our research focuses on the Traffic Signal Control problem since the efficient control of these intersections can significantly impact the productive hours and, through emission, the health of the citizens along with the depressing challenge of climate change. The Traffic Signal Control problem is well-studied and solved via several different techniques. However, most recently, Single and Multi-Agent Reinforcement Learning methods have arisen thanks to their performance and real-time applicability. Although rewarding schemes, which are the most crucial aspects of this method, do not seem to evolve at the same pace as the utilized techniques. In this paper, we propose a novel rewarding concept to compare its performance with the most common rewarding strategies in the literature. The results indicate that our approach outperforms its contenders from the literature in both classic and sustainability measures.

INDEX TERMS

Reinforcement Learning, Traffic Signal Control, Air pollution

I. INTRODUCTION

One of the 21st century’s most significant challenges that the entire society faces is global warming. This concept means a long-term change in the Earth’s climate system as a consequence of human activities. The primary factor is the burning of fossil fuels that increases the level of greenhouse gases in the atmosphere. The distribution of the different gases having relatively high Global Warming Potential (GWP) emitted in the U.S. in 2019 is shown in Fig. 1. It shows the ratio of methane in light brown, nitrogen oxides in yellow, carbon dioxide originated from fossil fuels in blue, other carbon dioxide emitted in dark brown, and other greenhouse gases in red color. One notable fact to highlight is that about 80% of the total amount is covered by only one component: carbon dioxide, making it one of the most influential contributors to climate change.

According to a 2020 survey carried out by the International Energy Agency, about 24% of the CO2 comes directly from the transport sector, from which road transport is responsible for over three quarters [1]. Another crucial problem, although closely linked to each other, is air pollution, in which road transport contribution has a similar share. Air pollutants, which are mostly end products of internal combustion engines, have significant effects not only on the environment but also on the human body.

There are already several approaches and solutions to mitigate these problems, such as the use of alternative energy-driven vehicles (e.g. electricity, hydrogen) [2], the application
of more efficient engines [3], the use of community car-sharing services, the use of alternative means of transportation and also the use of Intelligent Transportation Systems (ITS).

As many emission metrics are most affected by acceleration and its magnitude, it is advisable to intervene in locations where this parameter has a relatively high fluctuation – at the intersections. The fixed cycle traffic lights currently in use in several cities across the globe cannot adapt to traffic dynamics. Still, progress can be achieved by using different adaptive traffic control methods that operate based on the current traffic flow. Therefore, our research endeavor is to provide a solution that mitigates emissions and, along with that, increases the number of productive hours.

A. RELATED WORK
The control of signalized intersections is well-studied in single and multi intersection cases. The researchers utilized several different control approaches for these problems. Historically the Traffic Signal Control (TSC) problem is first solved with rule-based methods. The authors in [4] proposed a technique that utilizes filtering and fusion of traffic data to find platoons in the network. Furthermore, the filtering approach is used to predict the arrival time, and through this prediction, the signalized intersection can be controlled. In [5] the authors introduced a novel approach that can handle real-time traffic conditions and transit conditions as well. Multi-Agent Systems (MAS) are also a well-explored approach for solving the TSC problem in the case of complex networks with multiple intersections. In [6] the authors developed an algorithm that constructs signal groups and represents them as individual agents that operate on the intersection level in terms of decision-making. In their decisions, the agents always consider the given traffic scenario. Game Theory-based solutions are also a common approach in this realm. The authors in [7] formulated the control problem as a noncooperative game where the individual agents’ goal is to minimize some performance indicator monitored in their own part of the network. Many researchers have recently applied Dynamic Programming (DP) for the TSC problem, making it a well-explored approach. Most of the authors utilize these approaches thanks to their flexibility in terms of the traffic conditions where they can use them. Moreover, the tremendous versatility in potential performance indicators makes these algorithms tempting for such a control task. The authors in [8] proposed a novel Approximate Dynamic Programming based method that dramatically decreased the delay of the vehicles. Simulation-based algorithms are also serious contenders for solving the TSC problem. In [9] the researchers developed an algorithm that uses communication between the traffic signals and the sensors that monitor the traffic, and this method is successfully applied for a complex real-world network in a traffic responsive manner. Most recent approaches in this field exploited the potential of V2X technologies by the cooperation between the vehicles and the infrastructure to reduce delay and CO2 emission. The authors in [10] proposed a method that optimizes the signal phases and vehicle trajectories too and reaches a significant reduction of delay and CO2 emission. [11] proposes a cooperative method that optimizes the signal phases and the vehicles’ speed and successfully improves transportation efficiency and fuel economy. The authors in [12] proposed a Mixed Integer Linear Programming (MILP) approach for the control of Connected Automated Vehicles (CAV) for an isolated urban intersection and showed that it outperforms modern actuated control techniques. The previously introduced methods provide impressive results in solving the different forms of the TSC problem. However, as the complexity and size of the control problem increase, classical approaches lack the required computational resources to provide a real-time solution. Consequently, Reinforcement Learning-based method is living its renaissance in this domain. Thanks to their tremendous performance in this field and several transportation-related domains [13], [14], and also their potential for real-time applicability and solving complex sequential decision-making problems. In [15] the authors developed a state-representation. The core of the idea is the utilization of Convolutional Neural Networks (CNN) for this problem by dividing the lanes into cells and interpreting the cells as pixel-values. The number of channels is determined by choosing the measures representing a single vehicle in the given cell. This representation profoundly enhanced the performance of the same algorithms compared to feature-based value vectors. [16] use event-based data in the state representation, and the evaluations show that it can outperform both fixed-time and actuated controllers. [17] proposes a novel logic that influences the discount factor that is used in the Bellman equation for target value calculation. The results show that it can outperform the conventional DQN algorithm and the fixed-time controller. The authors in [18] conducted a comparison for the single intersection TSC problem between Game Theory, Reinforcement Learning, and rule-based methods. The results show that both approaches can outperform the rule-based fixed-time controller. All the previously presented RL methods nearly utilize the same type of rewarding concepts, which trains the network to develop a behavior that directly minimizes or maximizes a performance indicator measured directly from the intersection. On the
contrary, our proposed reward strategy exploits the fundamental operation of the TSC problem, which enables further reduction of delay and emission metrics. This paper evaluates our novel approach’s potential and compares it with different rewarding concepts from the literature and baseline methods. For a more comprehensive study on the TSC problem, see [19], [20].

B. CONTRIBUTION

Regardless of the control technique, the objective function has a crucial role in the final performance of the method. Reinforcement Learning algorithms are even more exposed to that through the rewarding concept since this is the only guidance an agent has in understanding the environment dynamics. Moreover, one can observe that the utilized algorithms in the TSC problem are evolved along with RL, and after a short delay, every innovation has occurred in this field. Although, the same cannot be said about rewarding concepts, which are improving at a much slower pace than representations and algorithms. Due to that, the main focus of the contribution is not to demonstrate that RL can reach better performance in classic or sustainability measures which problems are already explored thoroughly by other authors [21], [22]. But to show that the proposed rewarding concept is superior to other approaches from the literature based on an exhausting evaluation in both classic and sustainability performance indicators. Consequently, the contributions of the paper are the following:

- This paper proposes a novel rewarding concept similar to our previous approach [23] with one fundamental difference. The calculation of the standard deviation-based reward is based on the queue length of the lanes instead of the occupancies. This approach exploits the TSC problem’s fundamentals. The results show that it enables further reduction of CO2 emission and average waiting time compared to other methods from the literature and our previous concept.
- The Policy Gradient algorithm is trained with four different reward strategies. One is the proposed approach, and the other three are the most popular concepts from the literature.
- The trained agents are evaluated strictly under the same circumstances to seek representativity, both in classical and sustainability measures. In which cases the results demonstrate that the proposed approach has superior performance.
- The trained agents are also compared with the simple rule-based approach where the phase lengths are fixed, and a more advanced actuated Time-delay-based method, a built-in controller of the utilized SUMO software. The results show that the proposed method outperforms these control techniques in every measure in these cases.

II. METHODOLOGY

A. REINFORCEMENT LEARNING

Reinforcement Learning has earned the attention of researchers in the recent decade, thanks to its tremendous results in several competitive domains, such as videogames [24], robotics [25], and board games [26] as well. Thanks to these results, RL has always been considered a serious contender for providing the best solution in the TSC domain. The frequent use of RL in the TSC problem is the outstanding versatility in the field of sequential decision-making problems. One of these is the real-time applicability, which is due to the negligible decision-making delay caused by the neural network. In addition, non-linear function approximation supports scalability, which is an important feature, allowing to extend the solutions for a more complex network of intersections. Finally, it has a high generalization capability, which has been well-known and deeply researched from the beginning. The ultimate advantage of Reinforcement Learning over other Machine Learning methods is that it utilizes a reward signal to reach the desired behavior. Thanks to that, the accessible performance level is not limited by design but from the researcher’s intuitions who shape the rewarding concept and the state representation since these are the two major components that influence the credit assignment process.

The RL framework is formulated as the well-known Markov Decision Process (MDP) $< S, A, R, P >$, where the decision-maker is called an agent. The learning process is a series of interactions between the agent and the environment used to generate training data. Therefore, the function of the environment is twofold: firstly, it has to provide information for the agent, from which the decision-making is feasible, namely the state representation. Secondly, it needs to evaluate the consequences of the action chosen by the agent by providing a scalar value. This is the process of state transition and rewarding. The reward value is used to qualify an action. The training loop representing the above-described method can be seen in Fig. 2. During the training process, the agent’s Artificial Neural Network’s inner parameters are tuned using a loss function and the backpropagation algorithm. Two parts determine this function: the agent’s decisions and the feedback received from the environment.

1) Policy Gradient

One of the main algorithm families of RL is the so-called policy-based methods, which is of particular interest nowadays due to its emergent results in the field of control problems. One of the first algorithms of this family is the Policy Gradient algorithm along with Actor Critic methods which has the same basis but combined with value-based approaches. One significant advantage of policy-based methods compared to value-based methods is the guaranteed convergence to at least a local optimum [27]. In the case of PG, we tune the weights of the Neural Network to estimate the choice probability of each action as accurately as possible in every given state. Consequently, the output of the Neural Network is a probability distribution over the available actions from that specific state, which is determined by the
\[ J(\pi_\theta) = \sum_{t=0}^{T} r_t \cdot \log_2(P(a_i|\pi_i)) \]  

where  
\[ a_i \in A, i \in t = \{0; 1; 2; \ldots; T\} \]  
\[ r_t \]  is the reward received at the \( t \)th timestep by applying action \( a \) from state \( s \in S \).

Furthermore, the update of the \( \theta \) parameters of the neural network in case of PG is carried out according to (2):

\[ \theta \leftarrow \theta + \alpha \nabla \log \pi_\theta(s_t, a_t) \sum_{t=0}^{T} \gamma^t r_t \]  

where \( \theta \) denotes the weight matrix of the Neural Network, \( \alpha \) is the learning rate, \( r_t \) is the reward received at the \( t \)th timestep, \( \gamma \) is the discount-factor and \( \pi_\theta(s_t, a_t) \) is the probability of choosing action \( a \) from state \( s \) at the \( t \)th timestep.

### III. Environment

The environment has an essential role in the RL framework, notably interacting with the agent to generate training data for the training process. In our case, this task is performed by an intersection created in a simulation software called Simulation of Urban MOBility (SUMO) [28]. SUMO has been a serious contender for such a role since it supports real and artificially created traffic networks. It can monitor all meaningful traffic measures, which helps shape the state representation and the reward signal. Last but not least, it is possible to randomize the traffic flow, which enables the diverse gathering of training data. As can be seen in Fig. 2, our infrastructure operates as follows: the RL agent is implemented in python, with the application of Pytorch Deep Learning library. In the case of the environment, we can differentiate two separate components: a python component using the TraCI interface for communication and the transport network itself.

The geometric design of the network is shown in Fig. 3: four road segments meet at an intersection. Each segment is 500 meters in length, bi-directional, and 3.2 meters wide. At the end of each segment is a designated spawn point for the vehicles entering the simulation, which are randomly generated at each step. The intersection is controlled by traffic lights, and the agent’s task is to provide the optimal phase sequence. For evaluation purposes, we use the built-in function of the simulator, which monitors the impact of road traffic on the environment.

In each episode, for each entry point, we set the probability of spawning a vehicle to that specific point. Thus if we set this value to 1, one vehicle is generated at the given entry point in each time step. The randomization of the traffic situations can be managed in a way, that the environment is able to provide a variety of experiences to form a robust outcome.
A. STATE REPRESENTATION

State representation is one of the most crucial components of the RL abstractions since this is the only information an agent has about the environment. For the agent to understand its control task efficiently, it has to be given all the essential information that describes the dynamics of the environment. Of course, it is not trivial which measures to use in the representation. Hence it is solely the researcher’s responsibility the choose the right ones. Consequently, the reachable performance is exposed to how well the researcher understands the particular control problem. In the TSC problem, traditionally, there are two main approaches for state representation. The first is image-like representations such as raw RGB images and Discrete Traffic State Encoding (DTSE). The second is feature-based value vectors, where each lane is represented with a number of values.

In this paper, we utilize the second approach mainly because this information can be easily gathered and does not generate additional barriers between simulation and real-world application. In our case, a single value per lane has to be assigned, which in our case is the number of vehicles halting along the given road section. A vehicle is halting when the current speed is less than 0.1 \([m/s]\). The number of halting vehicles is divided by the maximum capacity of the given lane. Consequently, the state representation is a vector filled with four information, percentages describing the number of halting vehicles in the particular lane. This approach has another advantage: it can be extended to multi-lane scenarios with the following considerations. The halting vehicles in multiple neighboring lanes controlled by the same traffic lights can be summarized and divided by the summarized capacity of the lanes. While the state representation is extended with additional information if there is a neighboring lane controlled by a traffic signal with different green phases. These values are converted to the \([0,1]\) interval to avoid exploding and vanishing gradients during backpropagation.

B. ACTION SPACE

The appropriate choice of action space is also important along with the state representation since it fundamentally determines the level of generalization that can be reached due to the training process. The core concept is that the actions must have the same effect regardless of the particular environment state. According to the literature related to the TSC problem, two main approaches of the action-space can be encountered. In the case of the first one, after each phase, the agent has the ability to change the upcoming phases in the sequence, thus constructing the most promising one at the time. In the second one, the agent changes phase lengths by using continuous action-space with predetermined time-interval. In this paper, the agent uses the first approach with two phases that can be chosen:
  - Horizontal green is the phase where the horizontal lanes have green signals, while others have red.
  - Vertical green is the opposite when the horizontal lanes have red signals.

The length of the phases is fixed; hence the agent can switch the phase after a limited time, and it is crucial to state that there is no limit in how many times a phase can be chosen repetitively. Moreover, it is important to note that phase changes are separated with yellow and yellow-green signals. As the random traffic flow generation results in different loads on each road section, the agent must find the optimal time for switching the phases to get the best performance.

C. REWARD

The rewarding method is considered one of the main parts of the RL framework because these are the only feedbacks through which the agent can realize its actions’ consequences. Thus, the proper choice of rewarding can determine the success of the training. Since the rewarding concept’s formulation is non-trivial, we arrive at the same problem as in the state representation notably, that the researcher’s intuition influences the reachable performance. The rewarding concepts in the TSC problem do not evolve greatly in recent years. The authors mostly utilized direct measures of the given traffic scenario, such as waiting time, queue length, appropriateness of green time, average speed, and cost of phase transition \([29, 30, 31]\). The common misdesign called reward shaping \([32]\) is easily noticeable in these rewards. Since they focus on how the agent should behave rather than what we want to achieve in general, all these measures are only parts of what we want to achieve, and they can not concentrate on a single objective. Consequently, agents trained with these rewards will always be lacking in performance. In this paper, we propose a novel rewarding concept that exploits the inner dynamics of the TSC problem along with the proper formulation of the entire optimization task for the agent. The basis of our novel approach is the result of our previous research \([23]\). But contrary to that research, we utilize different features in the state-representation, notably the queue length scaled to the \([0-1]\) interval. But still, we interpret it as a mini distribution, and the goal of the agent is to minimize the standard deviation of this distribution. The main difference between this new concept and the previous one is that the simple lane occupancy is changed to the queue length. Although the rewarding formula we used has shown outstanding results against baseline controllers, the concept, in fact, has some performance limitations. The inertia of the traffic can explain this: when using occupancy, we do not have information about the kinematic parameters of the vehicles. So the resulting agent’s performance will involve an error occurring due to the inertia of the traffic. For example, if the occupancies of the lanes are equal, they will receive the same reward values at that time step, regardless of whether the vehicles in the lanes are stationary or in motion. This problem can be eliminated by changing the occupancy of the lane to the queue length, which is also interpreted as a distribution. So in our case, the reward of each step is calculated from the distribution of queue lengths normalized to the \([-1; 1]\) interval. In this case, a very close relationship is established between the three applied abstractions (state
representation, action, reward), since the state representation is the distribution of queue lengths, we can manipulate the mean and standard deviation of the distribution through the actions, and the agent’s task is to minimize the standard deviation. Therefore, we present an optimization task that is easy for the RL agent to understand, which is expected to result in fast and stable convergence, and extensive generalization capability.

D. TRAINING
One of the main aspects of the training process is the generation of sufficient but on the other hand, rather diverse traffic situations. Only in this way can we achieve the best possible performance, and generalization capability. In our case, it is resolved by using a randomly defined load on the incoming lanes. Therefore, the decision-maker is able to get a diverse set of experiences that can result in an extreme level of generalization and relatively quick convergence.

Each training episode begins with a warm-up stage when vehicles begin entering the network. It lasts until the global occupancy of the network reaches 10%. During the warm-up period, all traffic lights are given a red signal. Once the warm-up phase is over, the agent starts to control the traffic lights at its discretion with the actions described above. One training episode comes to its end if all the vehicles entered the network crosses the intersection. If the agent is unable to coordinate all vehicles participating in the simulation through the network, the episode will stop at the 1800th time step.

E. BASELINE CONTROLLERS
In the literature, most of the papers dealing with the TSC problem through Reinforcement Learning only evaluate their agents’ performance by comparing it to fixed cycle controllers as baselines or actuated controllers, and sometimes with other control approaches such as Game Theory and so forth. As the overall performance of the different solutions is getting better and better, the use of a fixed cycle controller is only reasonable for simplicity and ease of comparability due to its limited relatively low performance.

Therefore in this paper, the agent using the previously introduced rewarding concept (see Section III) is being compared with both a fixed cycle controller and a time-delay based actuated controller as baselines. The applied actuated controller is a built-in algorithm of the SUMO simulation software, which continuously monitors the incoming traffic, considers each vehicle’s accumulated time delays, and changes the traffic lights according to the time delays. For more information, see [28]. Along with the baseline controllers, we compare our method to the most commonly used rewarding concept in the literature (average speed) and the most frequently used rewarding concept that has been improved (average speed of occupancy).

As explained earlier (see Section I), the transport sector is a huge contributor to many problems concerning our planet. Thus for the sake of sustainability, any infrastructure development that is able to reduce environmental footprint is worth a further investigation. For this reason, the comparison of the solutions we present is demonstrated not only through the commonly used, classic parameters, but also with great emphasis on the environmental impacts of the methods, such as fuel consumption and the emission of carbon dioxide.

During the comparison phase, the performance of each solution is being measured over 1000 consecutive test episodes by generating the same incoming traffic flows with pseudo-randomness. A test episode lasts until the last vehicle leaves the transportation network.

IV. RESULTS
A. CLASSIC PERFORMANCE METRICS
Firstly, the methods will be compared in terms of classic parameters, which are defined as follows.

Average travel time is the sum of the total time spent in the transport network divided by the number of vehicles entering the simulation. It is formally described by (3) as:

\[ \bar{t}_t = \frac{1}{N} \sum_{i=0}^{N} (t_{i,\text{start}} - t_{i,\text{end}}) \quad (3) \]

The average waiting time is calculated as the sum of each vehicle’s accumulated waiting time divided by the number of vehicles. A vehicle is in halting state when its speed is less than 0.1 \( \frac{\text{m}}{\text{s}} \). For its mathematical meaning, see (4):

\[ \bar{t}_w = \frac{1}{N} \sum_{i=0}^{N} t_{i,w} \quad (4) \]

Queue length for each lane is calculated as the number of halting vehicles in the given lane. The queue length used for statistics is the sum of queue lengths for each lane, divided by the number of lanes. This result in a single value representing the total queue length in the network and is calculated according to (5):

\[ q_l = \frac{\sum_{i=0}^{L} q_{l,i}}{L} \quad (5) \]

The results displayed in Table 1 show that with the new rewarding, the performance of the PG agent outperforms SUMO’s built-in algorithm, fixed cycle controller, and all other rewarding concepts we have tested in terms of waiting.
Table 1. Statistical comparison of classical metrics

| Agents            | Avg. travel time [s] | Avg. waiting time [s] | Queue length [num. vehicles] |
|-------------------|----------------------|-----------------------|-----------------------------|
| PG–queue length   | 125.6 ± 23           | 31.5 ± 12             | 3.95 ± 9                    |
| PG–queue length std. | 154.7 ± 23           | 213.3 ± 7             | 3.61 ± 9                    |
| PG–average speed  | 161.7 ± 36           | 38.4 ± 23             | 5.62 ± 9                    |
| PG–occupancy std. | 155.1 ± 23           | 27.8 ± 9              | 4.34 ± 6                    |
| SUMO actuated     | 160.0 ± 25           | 47.3 ± 21             | 7.51 ± 14                   |
| Fixed cycle       | 384.4 ± 69           | 104.3 ± 22            | 8.88 ± 7                    |

Table 2. Statistical comparison of sustainability measures

| Agents            | CO2 [kg] | CO [g] | NOx [g] | PM10 [g] | HC [g] | Fuel [l] |
|-------------------|----------|--------|---------|----------|--------|----------|
| PG–queue length   | 110.1    | 3.5    | 48.0    | 2.4      | 18.7   | 47.5     |
| PG–queue len. std.| 108.5    | 3.3    | 47.2    | 2.3      | 17.7   | 46.6     |
| PG–average speed  | 118.7    | 4.1    | 52.0    | 2.6      | 21.3   | 51.0     |
| PG–occupancy std. | 110.2    | 3.5    | 48.1    | 2.4      | 18.7   | 47.4     |
| SUMO actuated     | 123.3    | 4.4    | 54.1    | 2.7      | 23.0   | 53.0     |
| Fixed cycle       | 229.5    | 10.8   | 102.5   | 5.2      | 54.6   | 98.7     |

FIGURE 4. Performance of the different solutions during the 1000.episode evaluation process

B. SUSTAINABILITY MEASURES

Table 2 shows the performance of different solutions to the TSC problem by comparing multiple emission parameters. The values are calculated by the SUMO simulation software per time step, thus only the averages of the values are displayed. The mathematical model used by SUMO is the Handbook Emission Factors for Road Transport, which calculates vehicle emissions based on polynomials derived from speed and acceleration. For more information [33].

It can be seen that the agent trained with the new rewarding formula outperforms not only the fixed cycle controller and SUMO’s built-in algorithm but also the most commonly used rewarding technique in the literature. As mentioned earlier, one of the most significant emission metrics related to global warming is carbon dioxide. Thanks to our improved rewarding concept, an additional CO2-emission reduction of 1.8% could be achieved compared to the original method, 12% compared to the SUMO algorithm, and 1.2% compared to the most common concept. These figures prove that the rewarding concept we proposed can effectively control the traffic lights at the intersection in terms of sustainability. In addition, it achieves extreme, 53% lower emissions compared to fixed cycle control, which is the most widely used control technology in the world.

Another critical metric is fuel consumption during the evaluation episodes. With our solution, we managed to further reduce fuel consumption. We improved two additional percent compared to the original occupancy-based rewarding and achieved 1.5% better results than the most popular rewarding. While this amount may not seem too much, the U.S. Energy Agency estimated, that in 2020, about 14.5 billion liters of fuel per day was sold globally. From this figure, it is clearly visible, that this 1.5% can save up to millions of liters in 24 hours, which can have a significant impact on the environment in the long run.

V. CONCLUSIONS

The agent’s task in the Traffic Signal Control Problem is to find the phase sequence combination that optimizes the traffic lights in terms of sustainability while reducing waiting- and travel time and the queue lengths for each vehicle at the intersection all at the same time.

Advanced adaptive control techniques can certainly reduce emissions in urban areas. Moreover, they also mitigate the
amount of time spent in traffic jams. In addition, the infrastructure (loop detectors) required for the state representation of the agent we introduced is already available in most major cities. Still, the ease of installation and reliability contribute to this specific solution’s wide applicability.

This paper first investigates the performance of the most common reward strategies applied for the TSC problem to compare with the concept proposed in this. Furthermore, we evaluated their performance based on classic measures such as average waiting time, average travel time, and sustainability measures, e.g., CO2 emission and fuel consumption. A Policy Gradient agent has been trained to apply four different rewarding concepts to solve the presented control task.

The results showed that with the application of our novel rewarding concept, the agent could achieve the best performance in classical parameters and sustainability measures except for one relatively small deviation compared to the approaches utilized in the literature. Furthermore, the agent using the newly proposed rewarding method outperformed both SUMO’s built-in actuated controller and the fixed cycle controller in sustainability measures and classical parameters.

In our future endeavors, based on the potential of the developed rewarding concept, our goal is to extend the approach to much more complex traffic networks and real-world networks to demonstrate how the proposed approach tackles these challenges. We will reformulate the training as a Multi-Agent Reinforcement Learning (MARL) problem since it has additional features that may enable us to reach superior performance, such as communication and even negotiations between the agents, hence between the intersections.

REFERENCES

[1] L. Al-Ghussain, “Global warming: review on driving forces and mitigation,” Environmental Progress & Sustainable Energy, vol. 38, no. 1, pp. 13–21, 2019.

[2] C. Acar and I. Dincer, “The potential role of hydrogen as a sustainable transportation fuel to combat global warming,” International Journal of Hydrogen Energy, vol. 45, no. 5, pp. 3396–3406, 2020.

[3] R. Hannappel, “The impact of global warming on the automotive industry,” in AIP Conference Proceedings, vol. 1871, no. 1. AIP Publishing LLC, 2017, p. 060001.

[4] P. Dell’Olmo and P. B. Mirchandani, “REALBAND: an approach for real-time coordination of traffic flows on networks,” Transportation Research Record, no. 1494, pp. 106–116, jul 1995.

[5] W. Ekelia, T. Sayed, and M. E. Esawy, “Development of dynamic transit signal priority strategy,” Transportation research record, 20111, no. 1, pp. 1–9, 2009.

[6] J. Jin and X. Ma, “A group-based traffic signal control with adaptive learning ability,” Engineering applications of artificial intelligence, vol. 65, pp. 282–293, 2017.

[7] I. A. Villalobos, A. S. Poznyak, and A. M. Tamayo, “Urban traffic control problem: a game theory approach,” IFAC Proceedings Volumes, vol. 41, no. 2, pp. 7154–7159, 2008.

[8] C. Cai, C. K. Wong, and B. G. Heydecker, “Adaptive traffic signal control using approximate dynamic programming,” Transportation Research Part C: Emerging Technologies, vol. 17, no. 5, pp. 456–474, 2009.

[9] D. McKenney and T. White, “Distributed and adaptive traffic signal control within a realistic traffic simulation,” Engineering Applications of Artificial Intelligence, vol. 26, no. 1, pp. 573–583, 2013.

[10] Y. Feng, C. Yu, and X. H. Liu, “Spatiotemporal intersection control in a connected and automated vehicle environment,” Transportation Research Part C: Emerging Technologies, vol. 89, pp. 364–383, 2018.

[11] B. Xu, X. J. Ban, Y. Bian, W. Li, J. Wang, S. E. Li, and K. Li, “Cooperative method of traffic signal optimization and speed control of connected vehicles at isolated intersections,” IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 4, pp. 1390–1403, 2018.

[12] C. Yu, Y. Feng, H. X. Liu, W. Ma, and X. Yang, “Integrated optimization of traffic signals and vehicle trajectories at isolated urban intersections,” Transportation Research Part B: Methodological, vol. 112, pp. 89–118, 2018.

[13] B. R. Kiran, I. Soth, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogaman, and P. Pérez, “Deep reinforcement learning for autonomous driving: A survey,” IEEE Transactions on Intelligent Transportation Systems, 2021.

[14] A. Fekhar, S. Aradi, and T. Bécis, “Fast prototype framework for deep reinforcement learning-based trajectory planner,” Periodica Polytechnica Transportation Engineering, vol. 48, no. 4, pp. 307–312, 2020.

[15] W. Genders and S. Razavi, “Using a deep reinforcement learning agent for traffic signal control,” arXiv preprint arXiv:1611.01142, 2016.

[16] S. Wang, X. Xie, K. Huang, J. Zeng, and Z. Cai, “Deep reinforcement learning-based traffic signal control using high-resolution event-based data,” Entropy, vol. 21, no. 8, pp. 744, 2019.

[17] C.-H. Wan and M.-C. Hwang, “Value-based deep reinforcement learning for adaptive isolated signal control,” IET Intelligent Transport Systems, vol. 12, no. 9, pp. 1005–1010, 2018.

[18] J. Guo and I. Harmati, “Comparison of game theoretical strategy and reinforcement learning in traffic light control,” Periodica Polytechnica Transportation Engineering, vol. 48, no. 4, p. 313–319, 2020. [Online]. Available: https://p Ajournals.com/article/35465/15580071/10.13031/1092-6771.2018.1344406.

[19] H. Wei, G. Zheng, V. Gayah, and Z. Li, “Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation,” ACM SIGKDD Explorations Newsletter, vol. 22, no. 2, pp. 12–18, 2021.

[20] M. Eom and B.-I. Kim, “The traffic signal control problem for intersections: a review,” European transport research review, vol. 12, no. 1, pp. 1–20, 2020.

[21] V. Haydari, M. Zhang, C.-N. Chuah, and D. Ghosal, “Impact of deep rl-based traffic signal control on air quality,” in 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). IEEE, 2021, pp. 1–6.

[22] Z. Xu, Y. Cao, Y. Kang, and Z. Zhao, “Vehicle emission control on road with temporal traffic information using deep reinforcement learning,” IFAC-PapersOnLine, vol. 53, no. 2, pp. 14960–14965, 2020.

[23] B. Kóvári, L. Szőke, T. Bécis, S. Aradi, and P. Gáspár, “Traffic signal control via reinforcement learning for reducing global vehicle emission,” Sustainability, vol. 13, no. 20, pp. 11254, 2021.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., “Human-level control through deep reinforcement learning,” nature, vol. 518, no. 7540, pp. 529–533, 2015.

[25] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” arXiv preprint arXiv:1509.02971, 2015.

[26] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton et al., “Mastering the game of go without human knowledge,” nature, vol. 550, no. 7676, pp. 354–359, 2017.

[27] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in Advances in neural information processing systems, 2000, pp. 1057–1063.

[28] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, “Microscopic traffic simulation using sumo,” in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 2575–2582.

[29] K. T. K. Teo, K. B. Yeo, Y. K. Chiu, H. S. E. Chou, and M. K. Tan, “Agent-based traffic flow optimization at multiple signalized intersections,” in 2014 8th Asia Modelling Symposium. IEEE, 2014, pp. 21–26.

[30] S. Araghi, A. Khodabakhsh, and D. Creighton, “Q-learning method for controlling traffic signal phase time in a single intersection,” in 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). IEEE, 2013, pp. 1261–1265.

[31] J. Jin and X. Ma, “Adaptive group-based signal control using reinforcement learning with eligibility traces,” in 2015 IEEE 18th International Conference on Intelligent Transportation Systems. IEEE, 2015, pp. 2412–2417.

[32] W. B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, and P. Stone, “Reward (mis) design for autonomous driving,” arXiv preprint arXiv:2104.13906, 2021.
[33] M. Keller, S. Hausberger, C. Matzer, P. Wüthrich, and B. Notter, “Hbeفا version 3.3,” Background documentation, Berne, vol. 12, 2017.

BALINT KOVARI received the B.Sc. degree in vehicle engineering, and an M.Sc. in the Faculty of Autonomous Vehicle Control Engineering from Budapest University of Technology and Economics, Budapest, Hungary, in 2018 and 2020. He is currently pursuing his Ph.D. at Budapest University of Technology and Economics, Budapest, Hungary. His research interest includes artificial intelligence, Reinforcement learning, vehicle dynamics, and mechatronics.

BALINT PELENCZÉI is currently pursuing his B.Sc. degree in vehicle engineering at Budapest University of Technology and Economics, Budapest, Hungary. His research interest includes artificial intelligence, mechatronics, machine learning, and intelligent transportation systems.

SZILÁRD ARADI (M’14) received the M.Sc. degree in 2005 and Ph.D. in 2015 from the Budapest University of Technology and Economics, Budapest, Hungary, where he is currently working with the Department of Control for Transportation and Vehicle Systems.

Since 2021, he has been an Associate Professor at the Department of Control for Transportation and Vehicle Systems, Budapest University of Technology and Economics. His research interests include embedded systems, communication networks, vehicle mechatronics, and reinforcement learning. His research and industrial works have involved railway information systems, vehicle on-board networks, and vehicle control.

TAMAS BECSI (M’14) Tamás Bécsi received the M.Sc. and Ph.D. degrees from the Budapest University of Technology and Economics, Budapest, Hungary, in 2002 and 2008, respectively.

Since 2005, he has been an Assistant Lecturer and since 2014, he is an Associate Professor, at the Department of Control for Transportation and Vehicle Systems, Budapest University of Technology and Economics. His research interests include linear systems, embedded systems, traffic modeling, and simulation. His research and industrial works have involved railway information systems and vehicle control.