IEEE Open Journal of
Intelligent Transportation
Systems

Received 18 August 2021; revised 10 February 2022; accepted 7 March 2022. Date of publication 11 March 2022; date of current version 22 March 2022.

Digital Object Identifier 10.1109/OJITS.2022.3158688

Reinforcement Learning-Based Traffic Control: Mitigating the Adverse Impacts of Control Transitions

ROBERT ALMS[1], ARISTEIDIS NOULIS[2], EVANGELOS MINTSIS[3], LEONHARD LÜCKEN[4], AND PETER WAGNER[1,5]

[1]Institute of Transportation Systems, German Aerospace Center (DLR), 12489 Berlin, Germany

[2]ARRC, Technology Innovation Institute (TII), Abu Dhabi, UAE

[3]Hellenic Institute of Transport, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece

[4]ICBM, University of Oldenburg, 26111 Oldenburg, Germany

[5]Institute of Land and Sea Transport Systems, TU Berlin, 10587 Berlin, Germany

CORRESPONDING AUTHOR: R. ALMS (e-mail: robert.alms@dlr.de)

This work was supported in part by the EC Project TransAID under Grant 723390.

ABSTRACT An important aspect of automated driving is to handle situations where it fails or is not allowed in specific traffic situations. This case study explores means, by which control transitions in a mixed autonomy system can be organized in order to minimize their adverse impact on traffic flow. We assess a number of different approaches for a coordinated management of transitions, covering classic traffic management paradigms and AI-driven controls. We demonstrate that they yield excellent results when compared to a do-nothing scenario. This text further details a model for control transitions that is the basis for the simulation study presented. The results encourage the deployment of reinforcement learning on the control problem for a scenario with mandatory take-over requests.

INDEX TERMS Connected automated vehicles (CAV), reinforcement learning (RL), take-over request (ToR), traffic management (TM), transition of control (ToC).

NOMENCLATURE

AV	Automated vehicle
CAV	Connected automated vehicle
CV	Connected vehicle
LoD	Level of demand
MRM	Minimum risk manoeuvre
MV	Manual vehicle
MDP	Markov decision process
No-AD	No automated driving
RL	Reinforcement learning
RSI	Roadside infrastructure
TM	Traffic management
TMC	Traffic management center
ToC	Transition of control
ToR	Take-over request.

The review of this article was arranged by Associate Editor Jia Hu.

I. INTRODUCTION

THE TREND towards vehicle automation and connectivity between vehicles (V2V) or infrastructure (V2I) implies a need for traffic management approaches to deal with emerging complications in future mixed traffic situations. In such scenarios, connected automated vehicles (CAV) can be addressed individually by V2I technology, which differs considerably from classic traffic management tasks that organize large numbers of road users with standard control measures like, e.g., signal control or ramp metering. Based on this communication, CAVs hold the potential to pose as sensors and actuators within the traffic system simultaneously. This opens up prospects for novel traffic management schemes.

Moreover, with the progressing deployment of CAVs that provide state-of-the-art level-two functionalities (see SAE taxonomy for automated vehicles [1]), the traffic system

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

will gradually turn into a system of mixed autonomy. Such a system presents various challenges in terms of traffic efficiency and safety when human drivers and partly to fully automated vehicles (AV) share the same road space. In particular, automation disengagements are of concern, i.e., when a human driver has to operate as a fallback for a failed vehicle automation and needs to respond in a proper and timely manner to take back the driving task [2]. This safety critical process, a so-called downward *transition of control* (ToC) [3], is an increasingly important studied research topic, especially from the perspective of manufacturers on how to design respective takeover strategies in highly automated vehicles [4], [5], [6].

In contrast, the macroscopic effects on overall traffic, which even successful downward ToCs may induce when occurring frequently in certain traffic situations and areas are less investigated so far. Therefore, in this paper we consider the issue of such transition areas from a traffic management perspective with a specific focus on:

1) how to model, simulate and manage downward ToCs, detailing some of the related work of the EC project *TransAID*, and on,
2) how to design a traffic management control that mitigates the adverse effects of downward ToCs on traffic with the help of tools from artificial intelligence compared to a more traditional approach.

To the best of the authors' knowledge, there is only one other publication presenting a model for ToCs that conducted simulations on traffic performance [7], but with a rather limited scope on restricting the lane change behaviour. Thus, we present a case study based on a novel ToC model, that demonstrates possible outcomes in future scenarios of mixed autonomy when downward ToCs contribute detrimental to the traffic performance on a macroscopic level. Our study introduces several approaches on managing those adverse impacts.

We apply *reinforcement learning* (RL) which is a concept in machine learning that formalizes a control task in form of a *Markov Decision Process* (MDP) [8] to maximize a reward in a trial-and-error learning process. Reference [9] points out that the terminus RL is a class of solution methods in machine learning as well as a research field of these solutions that work well on a problem. In that sense, this work simply uses RL as a method to solve a control task for traffic management rather than researching the problem of RL itself. RL-based methods have been previously adopted to optimize traffic light performance [10], control ramp meters [11], enable eco-driving along signalized corridors [12], deliver personalized driving policies [13], or facilitate big-data driven intelligent traffic management [14]. However, the RL-based management of downward ToCs upstream of a no-automated-driving zone (No-AD zone) has not been targeted by RL approaches up to date.

The rest of the paper is organized as follows. In Section II we introduce the modeling of ToCs and discuss the outcome of respective simulation results. Further, we briefly review some related publications in the context of traffic management of CAVs with the application of RL. Section III describes the computer experimental setup. That is, we define the traffic control task and formulate the MDP for the RL experiment. Section IV presents the simulation results and provides a discussion of the obtained results. Finally, in Section V we present our conclusions from this study and point to future research directions.

II. RELATED WORK

A. TRANSITIONS OF CONTROL

ToCs constitute overarching processes that govern bi-directional shifts of authority between the driver and the AV. In case of downward ToCs, factors endogenous or exogenous to the AV may force vehicle automation to disengage and request driver's intervention for resuming AV's control. The signal (audio, visual, haptic or combination of the latter) from the vehicle automation side that notifies the driver for the need to re-engage in the primary driving tasks is defined as a take-over request (ToR). A successful downward ToC is completed as soon as the driver has re-engaged and his/her situational awareness and driving skills are fully restored. If the downward ToC is unsuccessful, namely the driver does not respond to ToR within the available lead time, the AV stops as safely as possible via a minimum risk manoeuvre (MRM). In case of upward ToCs, the driver hands over control to AV within its Operational Design Domain (ODD) via the activation of its automated driving systems. For example, the manufacturer Daimler recently announced that it will introduce a conditionally automated level-3 system in 2022, which deploys this depicted takeover strategy [15].

Failure from the driver's side to react in a timely manner to ToRs has been linked with fatal crashes in reports from both the U.S. National Highway Traffic Safety Administration (NHTSA) and the National Transportation Safety Board (NTSB) [16]. Given the adverse impacts of downward ToCs on safety, multiple studies have ventured to identify contributing factors to automated vehicle disengagements, by harnessing data collected for the disengagement and AV collision reports of the California Department of Motor Vehicles [17]–[19]. Findings from the latter studies indicate that vehicle-initiated disengagements are positively correlated with sensing and planning issues on the vehicle side and occur with increased frequency in high speed driving conditions or when behavior from other traffic participants becomes unpredictable and irregular.

Another research branch on downward ToCs has placed emphasis on determining human and AV system factors that affect driving performance during the ToC-preparation and post-ToC phases. References [20], [21] conducted literature review studies to identify factors that influence response time to ToRs (available lead time, involvement in secondary tasks, take-over request functionality etc.) and post-takeover vehicle control (braking-steering) in different traffic situations. Moreover, they reviewed existing models suitable for capturing the aforementioned artefacts of driver behavior prevailing

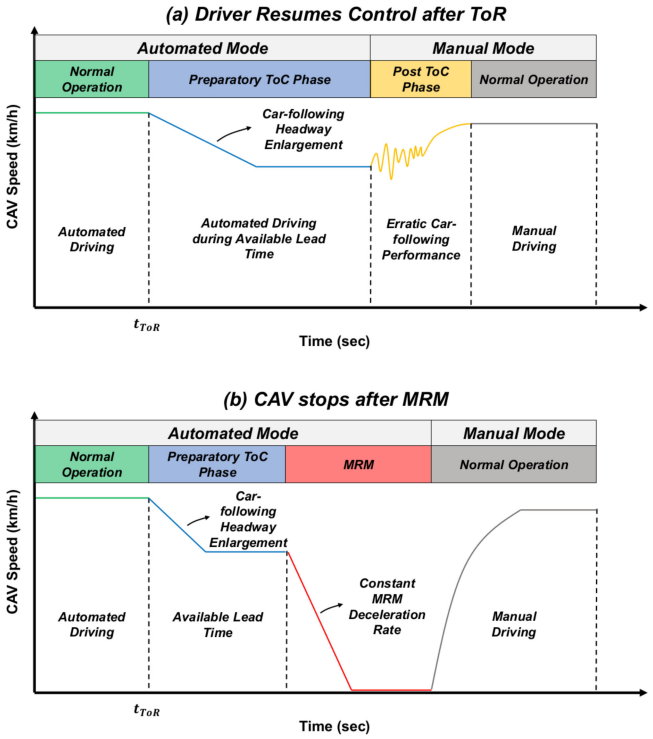


FIGURE 1. Illustration of the proposed ToC model for both cases: panel (a) shows a successful control transition; panel (b) presents an unsuccessful control transition resulting in a MRM.

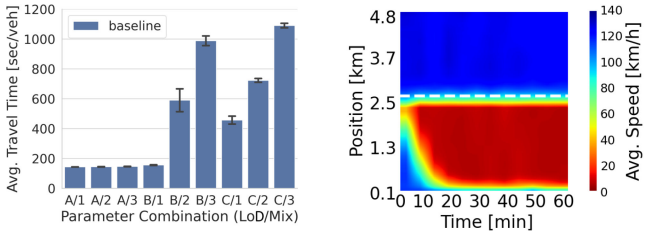
in the course of downward ToCs. High fidelity driver models that can comprehensively capture behavioural processes during downward ToCs are essential for exhaustively studying their impacts and assessing possible mitigation measures via computer simulations. To this end, [22] proposed a simulation framework that incorporated human factors (task demand and capacity, situational awareness) in microscopic traffic models which enabled the investigation of complex driver-vehicle interactions occurring during downward ToCs [23]. To capture the symptomatic side of potential disruptions of smooth vehicle operation induced by downward ToCs while, at the same time, enabling large scale simulations, [24] developed a simplified, computationally efficient model for ToCs, which allows capturing statistical characteristics of the take-over performance of AV drivers.

The modelling and large scale simulation of planned downward ToCs constitute focal elements in the context of this study. Since our primary interest lies on the statistical distributions of the downward ToC characteristics and associated potential disruptions of the smooth traffic flow, rather than on the detailed psycho-physical processes underlying these, we employ the ToC model adopted from [24]. A generic description of the ToC model is provided below, while its detailed mathematical formulations can be found in [24].

The proposed ToC model as illustrated in Fig. 1 is based on a state machine that enables transitions between automated and manual driving modes. ToRs issued during normal operation of automated mode prompt the commencement of a preparatory ToC phase when automated driving can be explicitly supported for a confined time interval (available lead time). Upon expiration of the available lead time, two distinct outcomes are possible according to driver's response to ToR. Either the driver reacts to the ToR and resumes vehicle control (Fig. 1, panel (a)), or the AV is forced to enter a minimum risk condition and stop as safely as possible (Fig. 1, panel (b)). In the context of the state machine, the first case pertains to the transition from the preparatory to the post-ToC phase, where the driver may exhibit a reduced driving performance until she/he fully restores her/his driving skills (normal operation in manual mode). The second case pertains to the execution of a minimum risk manoeuvre that safely stops the AV.

Our modelling approach encompasses the enforcement of lane change abstinence, acceleration abstinence, and the establishment of enlarged and secure car-following headways via a gap control mechanism throughout the preparatory ToC phase for safety reasons. The augmentation process of car-following headways can be manipulated with the adaptation of several calibration parameters of the latter mechanism (headway change rate, maximum allowed deceleration, duration of gap opening manoeuvre) to attain the desired car-following behavior (new desired headway, duration the new desired headway is maintained). Driving performance during the post-ToC phase is determined based on a driver state model that embeds perception errors in the default car-following behavior of the microscopic traffic simulator SUMO [25]. Each driver is randomly assigned an initial situational awareness state when she/he enters the post-ToC phase and a situational awareness recovery rate that regulates the restoration of situational awareness until normal operation in manual mode is achieved. Erratic car-following behavior is triggered during the post-ToC phase according to perception specific action points designed for imperfect driving [26]–[29]. Moreover, the considered ToC model assumes a constant deceleration rate during the MRM which can either take place in the vehicle's current lane or stop the AV on the right-most lane via lane changes (if surrounding traffic conditions permit).

Finally, the ToC model presumes two different ways for issuing ToRs in SUMO. In the first case, the location of ToRs can be a priori specified via a dedicated SUMO functionality. In the second case, ToRs are issued dynamically according to AV planned manouvres and surrounding traffic conditions. In specific, if an AV encounters a dead-end lane and is forced to execute a lane change for strategic reasons while nearby vehicles block the AV intended manoeuvre, then the AV will dynamically issue a ToR. On this occasion, the location of dynamic ToR becomes a function of AV speed and distance to the dead-end. Furthermore, dynamic downward ToCs encompass dynamical sampling of driver response time which entails probabilistic estimation of MRM frequency as well.



(a) Travel times for LoD A-C (b) Spatiotemporal speed diagram

FIGURE 2. **Baseline results of the TransAID use case study. Panel (a): Average travel times for different parameter combinations LoD/Mix, panel (b): Exemplary space-time-diagram for the mean speed of a single simulation run with LoD C and Mix 3. The white dashed line indicates the point from which all vehicles are obliged to drive manually.**

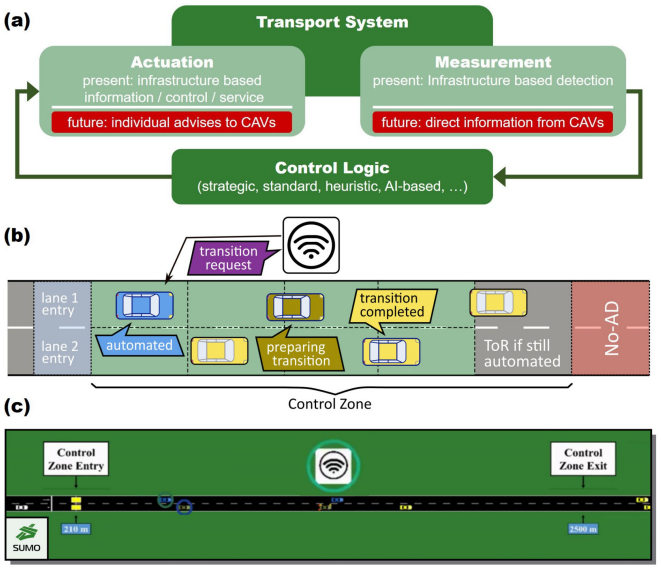


FIGURE 3. **Panel (a) shows the basic control scheme for a traffic management deploying CAVs for measurement and actuation in future traffic scenarios. Panel (b) illustrates a control scenario of a RSI communicating with automated vehicles within a defined control zone; colouring of vehicles corresponds to driving status; black dashed lines indicate the control cell borders. Panel (c) shows a SUMO screenshot of the discussed scenario.**

This approach for modelling ToCs makes it straightforward to embed the associated processes into microscopic traffic simulations of a broad variety of traffic scenarios at a high computational efficiency, such that the macroscopic assessment of impacts on traffic operations via microscopic simulation software can be achieved. A simulation analysis encompassing mixed fleet scenarios indicated that downward ToCs can induce adverse impacts on traffic efficiency, conflict risk and the environment in a variety of traffic situations (lane closure, road works, highway merge/diverge sections, no automated driving zones) [30]. Key findings of this study on the impact of downward ToCs on traffic flow upstream of a No-AD zone in the absence of any vehicle specific managing intervention from a Traffic Management Center (TMC)[1] are displayed in Fig. 2. In Panel (a) the average travel time of a single vehicle required to pass through the whole simulated road segment (cf. Fig. 3) is reported for an array of scenarios. These scenarios differ in the assumed level of demand (LoD), and the composition of the traffic, i.e., the percentage of automated vehicles. The LoD was varied ranging over the categories A, B, and C, and the percentage of different types of automated vehicles varied over Mix 1 (30%), Mix 2 (50%), and Mix 3 (80%), see Section III-C for details. Each scenario is assigned an ID composed of the corresponding LoD and Mix code, i.e., scenario C2 corresponds to LoD C and Mix 2. It can be observed that the do-nothing scenarios exhibit severe traffic jams beyond a certain demand level, clearly leading to a drop in speeds and a strong rise in the travel times. The space-time diagram in panel (b) shows the disruption of traffic flow caused by downward ToCs upstream of a No-AD zone in a single simulation for a specific parameter combination (LoD C, Mix 3).

To address the aforementioned impacts, [31] introduced several infrastructure-assisted traffic management measures designed for preventing, managing or distributing control transitions upstream of transition areas (areas on the roads where multiple control transitions may concurrently take place). Simulation findings showed that the proposed measures could mitigate the adverse impacts of control transitions for specific fleet mix and traffic demand scenarios. In particular, the distribution of downward ToCs in space and time upstream of a No-AD zone based on the TransAID approach could in many cases prevent traffic disruptions. The proposed distributed scheduling of downward ToCs reduced the local accumulation of accelerations/decelerations which may lead to strong speed variations and can consequently generate unsafe conditions. Thus, the rise in travel times observed in Fig. 2 could be at least postponed to a higher demand, or, in the best case, be avoided at all by AV specific ToR scheduling. Subsequently, the management of downward ToCs and guidance of MRMs to safe spots has been explored via real-world testing by Coll-Perales *et al.* in [32]. Their results suggested that the provision of personalized advice to CAVs including information about recommended ToC and safe spot locations could minimize MRMs taking place in lane which can result in hazardous situations.

B. REINFORCEMENT LEARNING IN TRAFFIC CONTROL

RL is a branch in machine learning that is heavily researched in emerging trends of deep learning applications nowadays, partly inspired by recent influential publications from [33] and [34]. Intelligent transportation systems (ITS) with often highly complex control tasks and vast volumes of data are a particularly promising field for data-driven techniques to improve nonlinear models as well as developing novel ideas to tackle future transportation challenges. The deployment of RL in traffic research was propelled by its ready utilization on control tasks in connection with traffic simulators. There is a wide range of traffic simulations that emulate a

1. CAVs will be informed about the existence of the upcoming No-AD zone via a simple information message, which causes automated vehicles to hand over control to the human driver at a specific position.

certain system behavior. These representations of dynamic traffic systems conveniently serve as the environment component in RL. The feasibility to train an agent that operates and handles specific tasks in such environments allowed the rapid testing and progress of RL related techniques in traffic control. As [35] point out in their survey, various traditional transportation problems like demand or destination prediction, travel time estimation, traffic signal control or traffic flow prediction are investigated with help of deep learning and RL techniques.

Other areas of research and application where deep learning and RL have proven useful are traffic signal control [36], [37], [38] connected automated vehicles in mixed autonomy traffic [39], variable speed limit control at bottlenecks and ramps [40], or at round-abouts [41], to name but a few.

III. EXPERIMENTS

A. COMPUTER EXPERIMENTAL SETUP

We define the main control task as follows: A Traffic Management Center has to address vehicles within a confined area to prevent them from entering a No-AD zone,[2] still driving in automated mode. Therefore, a generic two-lane road is divided into two regions: (1) an upstream area denominated the control zone where automated driving is allowed and (2) a downstream area, where manual driving is mandatory (No-AD zone). Thus, the TMC issues a ToR to every CV and CAV approaching the No-AD zone. Vehicles that receive a ToR, initiate a downward ToC. A downward ToC results, for a certain time-span, in a moderate vehicle deceleration caused by the gap enlargement in the ToC preparation phase, and possibly reduced human driver performance. The control zone is managed by the TMC via V2X communication, e.g., using roadside infrastructure (RSI). The RSI receives position, speed and driving status from each CV and CAV via cooperative awareness messages [42] and maneuver coordination messages [43].

Fig. 3, panel (a) illustrates the envisioned role of connected vehicles within the standard control scheme of traffic management, actively adding sensory information and direct actuation to the control loop. Panel (b) details this simple idea of an RSI communicating with vehicles within the control zone by sending individual ToRs, aiming to prevent those vehicles to exit the control zone while still driving in automated mode. Panel (c) shows a snapshot of such a simulation experiment in SUMO.

For simplicity, failing downward ToCs resulting in MRMs are not managed specifically in this scenario.[3] Although MRMs are included in the simulations with a very rare average rate of occurrence, we do assume that human drivers resume the vehicle's operation timely and do not block one lane, or both lanes, for an extended period of time. In the majority of the events, the driver takes over before the vehicle has stopped completely. Thus, occurring MRMs still have a negative impact on the traffic flow, but from a macroscopic perspective, they play only a secondary role relative to the by far more numerous downward ToCs. Fig. 4 shows the respective probability distribution of the reaction times by human drivers to take back control of the vehicle after receiving a ToR.

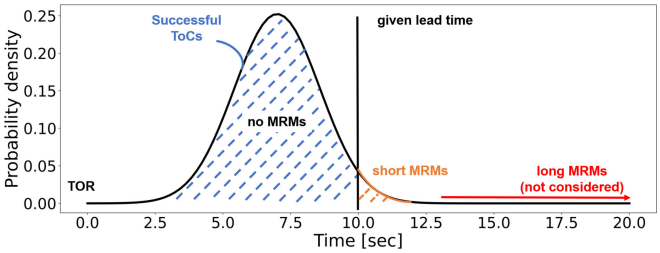


FIGURE 4. Reaction time probability distribution of the scenario.

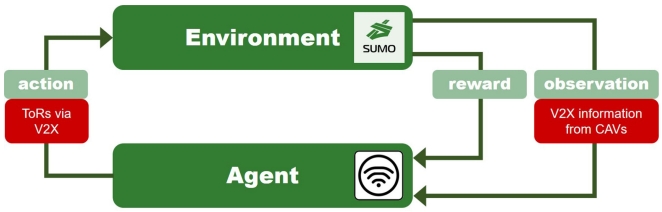


FIGURE 5. Interaction between agent and environment in a MDP (green boxes) including the prospective role of connected automated vehicles in this context (red boxes).

B. FORMULATION OF MARKOV DECISION PROCESS (MDP)

An MDP is a formalization of the decision making process of an agent that interacts with its environment. Specific actions by the agent affect the state of the environment. Based on observing the current state of the environment, the agent can take actions that aim to maximize future rewards, where rewards represent desirable outcomes. Fig. 5 schematically shows such an interaction.

In reference to Fig. 3, the environment represents the transport system, in our case emulated with the traffic simulator SUMO, and the agent corresponds to the traffic control logic. The data obtained by measurement and the means of interaction via V2X technology provide a basis for defining specific actions and observations in the MDP in this context. RL uses this formalization in order to guide and improve the decision making of an agent in such an interactive environment.

In this work we train a policy with Twin-Delayed Deep Deterministic Policy Gradient (TD3) [45], using the implementation of the *stable-baselines* library (version 2.10) [46]. A TD3 algorithm, which is a modification of Deep Deterministic Policy Gradient (DDPG) [47], simultaneously learns a Q-function for value updates and a

2. In this study we do not further examine the rationale for the mere existence of a No-AD zone. We presume the No-AD zone as a prerequisite for the TMC to address automated vehicles not to enter this area while driving automatically.

3. We refer the interested reader to [44] for a concept of how MRMs could be guided to safe spots if the given infrastructure permits.

target-policy that maximizes the Q-function through gradient ascent. The interplay between those two is referred to as an actor-critic algorithm. TD3 improves on DDPG, by adding three tricks: (1) estimating the target with two Q-functions (clipped double Q-Learning), (2) employing a lower update rate on the target network (delayed policy update) and (3) adding random noise to the target policy (target policy smoothing) [47].

In the following, we formulate two models, *model 1* and *model 2*, which differ in observation space and reward function. The definitions for observations, actions and rewards in the next paragraphs are derived from initial work in [48].

1) OBSERVATIONS

The standard RL approach requires that the space of all possible states is of a fixed dimension. Therefore, the description of a dynamic vehicle flow as a list of vehicle states is not suitable, as this list greatly varies in length over time. We propose to overcome this problem by dividing the control zone of the highway into a constant number of cells per lane. In this manner, the agent can operate on a constant state space, despite facing a dynamic number of vehicles. The cells represent smaller parts of each lane with approximately the same size (cf. Fig. 3, black dashed lines indicate cell borders). We found that a number of $N_{\text{cells}} = 14$ cells represents a good trade-off between complexity and granularity for the scenario at hand. Every cell is described by three respectively four values, which constitute the perception of the environment for the agent:

(i) the average speed of all vehicles in the cell,
(ii) the number of manually driven vehicles (MVs) in the cell,
(iii) the number of CAVs in the cell,

For *model 1*, these quantities constitute the complete state of the environment. For *model 2*, we further provide

(iv) the number vehicles that are about to enter the control zone in the next time steps (see paragraph *c*) below for details).

2) ACTIONS

Similarly, each cell represents a potential target object for an action of the agent, i.e., the TMC. Effectively, an action would correspond to the transmission of ToRs to all CAVs in a specific cell at a specific time step. However, formally we construct the action in the MDP framework as a vector

$$\boldsymbol{a} \in [0, 1]^{N_{\text{cells}}},$$

which assigns a probability a_i, with $0 \leq a_i \leq 1$, to each cell, with which an effective action is triggered. Thus, at each control step (that is every second), the agent updates the probability for a ToR to every cell, taking into account the observed state. In this context, an assignment $a_i = 1$ corresponds to the deterministic decision of sending ToRs to cell i, while $a_i = 0$ ensures that no ToR is emitted.

Furthermore, the protocol ensures that a vehicle, which passes one of the last cells before the No-AD zone, will always receive a ToR, if it hasn't already. This is to avoid the vehicle from entering the No-AD zone while still driving automatically.

3) REWARD

At each control step, a reward

$$r(t) = r_{\text{ToR}}(t) + r_v(t), \tag{1}$$

is calculated based on a part $r_v(t)$ associated to the state of the environment and a part $r_{\text{ToR}}(t)$ associated to the number of effective ToRs transmitted by the TMC. For both considered models, we employ a ToR reward

$$\begin{aligned} r_{\text{ToR}}(t) = &\sum_{i=1}^{N_{\text{cells}}} r_{\text{ToR},i} \cdot \#\{\text{ToRs sent to cell } i\} \\ &- \pi_{\text{ToR}} \cdot \#\{\text{ToRs sent to last cells}\}. \end{aligned} \tag{2}$$

Here, $r_{\text{ToR},i}$ is a constant which defines the reward associated to a transmission of a ToR to a CAV in the i-th cell, which increases linearly from zero to w_{ToR} with the cell's index along its lane. That means, the further downstream a CAV is located when receiving its ToR, the higher the associated reward. As an exception, if the ToR is sent just before the No-AD zone, a penalty π_{ToR} is imposed, since a belated transmission increases the risk for CAVs to enter the No-AD zone in automated mode.

The fraction $r_v(t)$ of the reward (1) applied in *model 1* is proportional to the average speed $\bar{v}(t)$ taken over all vehicles in the control zone:

$$r_{v,\text{model1}}(t) = \bar{v}(t)/v_{\max}, \tag{3}$$

where $v_{\max} = 36m/s$ is the maximal allowed speed in the scenario.

In *model 2* this definition is extended by terms accounting for vehicles loaded into the simulation, but not entered, yet. For clarity, given a specific demand level, the simulation software SUMO generates vehicles at a corresponding rate and tries to insert these into the simulation scenario. If there is not sufficient free space on the road, it keeps the generated vehicles in a buffer. That means, these "pending" vehicles represent a tailback not depicted in the simulation and contain information regarding the traffic situation, valuable to its evaluation. Accordingly, we define the reward for *model 2* as

$$r_{v,\text{model2}}(t) = \bar{v}(t) + \bar{v}_{\text{pend}}(t) - \pi_{\text{pend}} \cdot \#\{\text{pending vehicles}\}, \tag{4}$$

where $\bar{v}_{\text{pend}}(t)$ is the average speed of all vehicles in the control zone *and* in the simulation buffer (accounted for with speed zero), and π_{pend} is a constant scaling the penalty imposed per loaded vehicle, not yet inserted in the simulation.

TABLE 1. Demand levels.

	Level of demand (LoD)		
	A	***B***	***C***
Q_{in} [veh/h]	1470	2310	3234

TABLE 2. Traffic compositions with vehicles shares for three different mixes.

Traffic mix	Vehicle Type		
	MV	***CV***	***CAV***
Mix 1	70%	15%	15%
Mix 2	50%	25%	25%
Mix 3	20%	40%	40%

TABLE 3. Vehicle types in the simulation represented by SUMO model combinations. The Krauß model is SUMO's standard model, while the ACC model is described in [49].

Driving Mode	SUMO Model	Vehicle Type		
		MV	***CV***	***CAV***
Car Following	Krauß	o	-	-
	ACC	-	o	o
Lane Change	Default	o	-	-
	Parametrized LC	-	o	o
Control Transition	ToC	-	o	o

C. SIMULATION AND TRAINING SETUP

We conducted our simulations with the microscopic traffic simulator SUMO [25], version *1.6*. On a two-lane motorway with a length of 5.0 km and a speed limit of 130 km/h, vehicles enter the network randomly with a Poissonian distribution with a demand Q_{in} at the upstream part of the road. The maximum capacity of the two-lane motorway is assumed to be 4200 $[veh/h]$ for a homogeneous fleet of MVs, that means without the existence of control transitions. Since the maximum road capacity inevitably decreases in the presence of control transitions in all cases, and also in every case differently, we compare those TM approaches based on the induced demand. Three different demand levels *LoD* were defined (see Table 1).

The following tables summarize the three different traffic mixes with shares of manually driven (MV), connected (CV), and connected automated vehicles (CAV) used in the experiment (see Table 2), and also the respective vehicle models represented by the different models in SUMO (see Table 3). The selection of the simulated traffic mixes was made to verify that increasing shares of CV/CAVs escalate traffic disruption due to higher frequency of downward ToC events which are not efficiently distributed in space and time, rather than to explicitly quantify the exact penetration rates of CVs/CAVs inducing traffic flow breakdown as a result of accumulated downward ToCs. In the context of our simulation experiments, we consider that drivers of CVs continuously monitor the operation of the automation functions and can promptly react to ToRs. Moreover, CVs are ACC/CACC capable and their lane change behavior is more conservative compared to MVs. On the other hand, drivers of CAVs can be involved in secondary tasks during normal operation in automated mode, and thus exhibit delayed response to ToRs or even fail to resume vehicle control within the available lead time (CAV executes MRM on these rare occasions). CAVs are also assumed ACC/CACC capable, but their lane change behavior is more conservative compared to CVs. A detailed parametrization of the utilized SUMO models per vehicle type can be found in [30].

In the simulation, the No-AD zone starts at 2.5 km downstream of the network entry. The TM controller addresses CVs and CAVs ahead of the No-AD zone by sending take-over requests via the SUMO API *traci*.

For the parametrization of the TD3 training we mainly use the default parameters provided by the *stable baselines* library. For approximating the policy and the Q-function we deploy neural networks with two hidden layers, using [300, 400] for layer size as in [47]. The only hyperparameters, which have been altered in favor of the best results, are:

1) *buffer_size*: The size of the replay buffer to save experiences at each step, to update the target network episodically was set to 100000.
2) *train_freq*: The value to define the model update rate of the target network was set to 300 steps.

Two different models termed *model 1* and *model 2* were trained with the respective rewards functions. For the respective reward parameters we chose: $w_{\text{ToR}} = 10$, $\pi_{\text{ToR}} = 100$, $\pi_{\text{pend}} = 1/1200$. A training episode ran for $1200s$, with the first $200s$ as uncontrolled warm-up in order to establish a fully populated control zone. We let the model learn for 2000 episodes with a random seed for each episode. Both policies were trained on the highest demand level *C* with the vehicle fleet of *Mix 3*.

After the finished training, we ran simulations with both models for all parameter combinations (LoD/Mix), each combination with 10 random runs for 1 hour simulated time per seed.

IV. RESULTS

In the following we present the results obtained from the simulation study. We ran simulations for all parameter combinations (LoD/Mix) for *model 1*, *model 2*, the do-nothing case (*baseline*), the deterministic control approach proposed by the *TransAID* consortium [31], and a *random* model, that encompasses uniformly distributed ToRs within the control zone. Since our RL-based models were explicitly designed with the consideration of mobility objectives, we focus on the analysis of simulation results from the traffic efficiency perspective. First, we evaluate the performance of all cases based on the average travel times per vehicle and the average distance driven in automated mode (CAV distance). Also, we compare the control strategies for both trained models based on space time diagrams and the spatial distribution of sent ToRs. Finally, we analyze the training success by evaluating the reward.

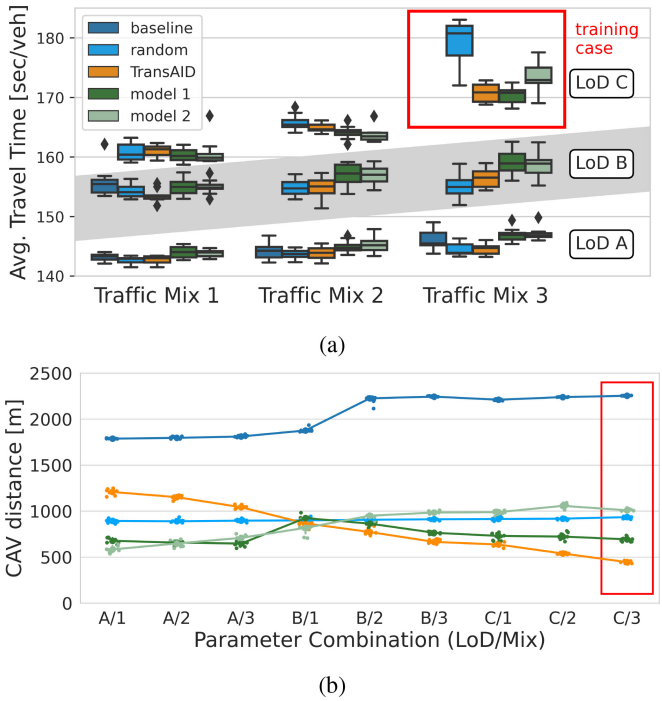


FIGURE 6. **Aggregated results for five different cases/models: *baseline* (dark blue), *random* (light blue) *TransAID* (orange), *model 1* (dark green), *model 2* (light green). Panel (a): Travel time for all parameter combinations; results of the baseline simulation for parameter combinations higher than B1/Mix 1 are excluded. The medians for those excluded combinations in ascending order are: 620.3, 978.2, 457.6, 721.8 and 1093.6*s*. Compare also to Fig. 2, panel (a). Panel (b): Average CAV distance driven in automated mode. The lines are drawn to guide the readers eyes.**

A. MOBILITY PERFORMANCE

Fig. 6 presents the aggregated results of the simulation study for all parameter combinations for both models compared to the unmanaged (baseline) and also two alternatively managed cases (TransAID) and (random). Panel (a) shows the individual vehicle's average travel times in the form of boxplots. Panel (b) respectively presents the average covered distance of CAVs within the control zone driving in automated mode.[4]

Firstly, comparing only the two trained models with each other, we observe that travel times are almost the same for all combinations except for C/3 where *model 1* slightly outperforms *model 2*. Notably, this is the combination the models were trained with (cf. red boxes in Fig. 6). On the other hand, for C/3 *model 2* covers significantly more CAV distance than *model 1* (about plus 300*m* per vehicle on average, which corresponds to approximately one cell length), corresponding to the rationale of the model development, see Section III-B. For parameter combinations C/2 to B/1 *model 2* is able to prolong automated driving for CAVs longer than *model 1*. For LoD A differences in CAV distance are rather small. So, the plain side-by-side comparison of both models leads to the initial conclusion that both achieve a good compromise between optimizing travel time and CAV distances simultaneously, yet with different emphasises on favouring travel time vs. CAV distance (cf. Fig. 6, case C/3).

The good performance of both models becomes obvious at higher demand levels when compared to the the other three approaches. For the baseline, we first take a look at panel (b) with the CAV distance. Due to the fact that all CAVs perform the control transition shortly before entering the No-AD zone, they cover the maximum possible CAV distance, but since consecutive and simultaneous downward ToCs cause disruptions in traffic flow, the travel time in panel (a) increases significantly (up to the factor 6, see Fig. 2) which ultimately results in traffic jams for parameter combinations higher than B/1.[5] In contrast, the original approach from TransAID is based on distributing individual ToRs, by forming virtual platoons for consecutive CAVs. This allows a continuous optimization of the latest possible ToR depending on the current speed of a CAV, achieving quite low average travel times, in fact slightly better than *model 2* and similar compared to *model 1* (see panel (a)). But, since this approach strongly depends on the traffic density within the control zone, in panel (b) we can observe a steady decrease in CAVs distance for higher parameter combinations. Both models prevent this decline for LoD B and LoD C, but drop noticeably in CAV distance for LoD A.

Comparing the random case to both models, it is noticeable, that this rather simple heuristic approach performs very well for LoD A and LoD B in terms of travel time and CAV distance, similar to *TransAID*. For C/3 though, this random approach apparently hits the capacity limit way earlier and shows significantly higher travel times.

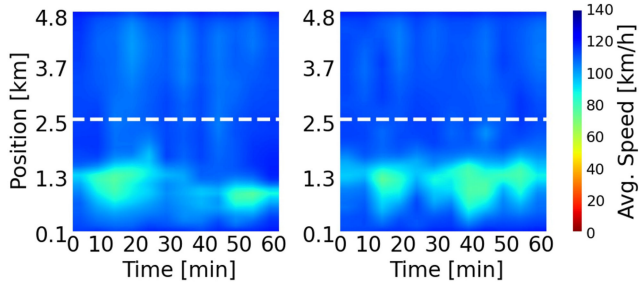
Overall, the highest benefits of the two RL models compared to all other cases can be gained at LoD C depending on the performance metric (travel time vs. CAV distance). Both models seem to be able to handle lower demands and mixes, i.e., LoD B/Mix 2 and Mix 3, that are somewhat close to the training case, but for LoD A that mobility performance drops for the RL models.

B. CONTROL STRATEGY PERFORMANCE

In Fig. 7 two exemplary space time diagrams are shown, which correspond to the same parameter combination as in Fig. 2. They illustrate that both models are very well able to distribute downward ToCs so that no traffic jams develop. The visible differences between panel (a) and (b) in the distribution of the mean speed within the control zone (the area below the white dashed line) result from the different control strategies of the two models. The areas of lighter blue indicate the occurrence of short episodes of slower

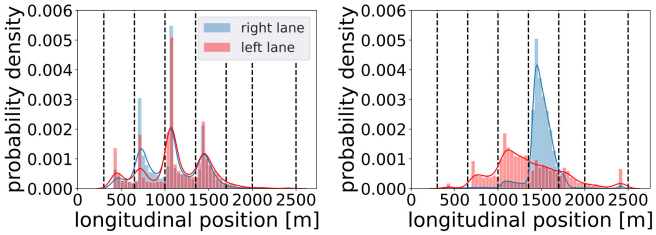
4. Note, for better visibility, in panel (a) we excluded boxes of the baseline simulation for parameter combinations higher than B1/Mix 1. In panel (b) we used grouped scatter plots and drew additional lines between groups for each model.

5. Notice that for the baseline, the average covered CAV distance for demand level LoD A and B1/Mix1 is less than 2000m opposed to the rest of the combinations with 2400m. This is because the ToR message is triggered dependent on the current vehicle speed. For combinations with low travel times, vehicles can drive almost up to their desired speed, which is about $36m/s$ opposed to the rest of the parameter combinations, where vehicles in congested traffic only drive about $5 - 10m/s$.



(a) *model 1* (b) *model 2*

FIGURE 7. Spatiotemporal diagrams of mean speed for a simulation run with *model 1* (left) and *model 2* (right) of a single simulation run: LoD C - Mix 3 - Seed 7. The white dashed line represents the point from which all vehicles have to drive manually.



(a) model 1 (b) model 2

FIGURE 8. Spatial probability distribution of take-over requests sent to CV/CAVs within the range of the control zone for *model 1* (left) and *model 2* (right) for parameter combination LoD C/Mix 3. The black dashed vertical lines mark the cell edges defined in Section III-B, showing 7 cells per lane.

average speed, which do not build up to a persistent congestion, though. Also, they support the argument that the TMCs objective is to preserve a rather smooth traffic flow without too many disruptive decelerations/accelerations by consecutive vehicles at similar positions. Clearly, *model 1* and *model 2* improve the smoothness of the traffic flow in comparison with the unmanaged case shown in Fig. 2. Thus, the rationale of the reward design seems to have identified mechanisms related to the emergence of congestion in that case. The trained controllers manage successfully and counteract congestion by a distribution of ToRs.

However, the spatiotemporal diagrams rather indirectly point to the control strategy, since the vehicle speed is a delayed indicator of the TMCs distribution strategy. Therefore Fig. 8 shows the spatial distribution of ToRs sent to CAVs within the control zone for each lane. As a clear distinction, we observe a disparate utilization of the left and right lane between both models, as well as different spatial distributions over the length of the control zone. Whereas *model 1* uses both lanes almost equally close to a normal distribution pattern around 1050*m*, *model 2* shows significant differences in distribution (skewed distribution on left lane) and lane utilization shifted further downstream (1600*m* on right lane). This might be connected to differences in the overall performance of the control strategies. On the contrary, both models show similarities in not sending ToRs to the last cells before the No-AD zone (longitudinal position >2000m), as imposed by the reward definitions.

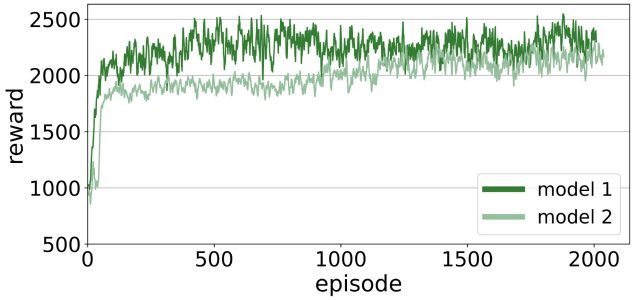


FIGURE 9. Reward per episode for *model 1* and *model 2*.

Despite these differences, the ToC management performs well in both cases while achieving a compromise between travel time optimization versus preserving the automated driving mode.

C. TRAINING ASSESSMENT

Fig. 9 shows the reward curves for the training process of the RL-based models 1 and 2. Both models converge relatively quickly, although *model 2* apparently needs more episodes to do so. Given that we effectively trained for 1000 steps per episode, the final average return is about 2.2 - 2.4 for both models. However, note that the average return for *model 2* may be expected to deviate slightly due to the different definition for $r_v(t)$ in Eq. (1), cf. (4).

Overall, some of the variance in the reward could be explained due to the randomness in vehicle flow and numbers. For the training of models 1 and 2, leading to the performance reported in Section IV, we ran each training run for 2000 episodes with random seeds per episode.

Additionally, we tested training runs with fixed seeds, i.e., deterministic vehicle flows for each episode, in order to find a smoother reward convergence (as e.g., in [50]). Those test runs (not displayed here) indeed showed less variance in the reward and also converged at about the same level as the training with random seeds. Nonetheless, the random seed training induced a better overall performance. We consider this to be attributable to an increased versatility of the control gained from the confrontation with a larger variety of situations and allowing it to handle different LoD and traffic mixes in a more robust fashion.

Moreover, it is worthwhile mentioning that the choice of the reward definition is critical for the training success, as well. For example, during the development, we tested one function, which imposed rather strong penalties (i.e., negative rewards) on the sending of ToRs. The penalties included a similar cost gradient as in (2), which favored longer distances in automated driving mode. From this definition, the model did not learn to postpone ToR transmissions, as intended. In contrast, it rather learned to reduce the occurrence of ToRs via a reduction of the inflow. This was achieved by sending all ToRs to the first cell in contradiction to the original objective of prolonging the distance of automated driving.

Clearly, an accurate design of the reward is necessary to align the resulting model behavior with the training objectives.

V. CONCLUSION

For a range of different demand levels and ratios of automated and manually driven vehicles, we have demonstrated, that traffic control can significantly attenuate the negative side-effect of control transitions from automated to manual. Four different TM methods have been explored and compared to a do-nothing scenario: Two heuristic approaches adhering to the form of conventional TM protocols, and two based on a RL approach employing slightly different reward functions. All methods perform similarly well, in terms of travel times and slight improvements with the trained RL models for prolonging the automated driving mode. Furthermore, all protocols outperform a do-nothing solution. Especially for higher demand levels the performance gap is significant. It is noteworthy that the RL models were trained with just one, relatively high demand at a fixed, relatively high ratio of automated vehicles in the traffic composition, but still perform properly when getting applied to lower demands and other fleet mixes in our scenario.

The robustness of the RL approaches has been tested by changing the demand and the vehicle fleet composition. In all cases, we see that the traffic management can ameliorate the potential capacity drop induced by an accumulation of downward ToCs in transition areas. We assess that this delay of the foreseeable capacity drop under high demand and with high CAVs percentages will be the task at hand for future TMCs in comparable scenarios. However, for very large demands close to capacity limit all of the methods must finally fail. While the two RL models, as well as the conventional TM controller, do not differ considerably in the maximum capacity they finally achieve, they enhance the capacity significantly compared to the baseline approach. The conventional TM controller *TransAID*, although way better than a do-nothing approach, manages high demand levels only by compromising the objective of preserving the automating driving mode. On the contrary, the naive random distribution TM approach achieves better CAV distances than *TransAID*, but only performs adequately for lower demands and hits the capacity limit earlier than the RL models.

We acknowledge that the sole objective to maintain automated driving as long as possible, although a downward ToC cannot be avoided in such a scenario, might be conflicting from a traffic management perspective and should only be considered while simultaneously tackling the adverse impacts discovered in the baseline analysis. Manufacturers and consumers of CAVs on the one hand might be interested in maintaining automated driving features without external interference, while from a traffic safety and efficiency perspective it can be favourable to initiate downward ToCs further upstream. In denser and highly heterogeneous traffic, belated ToRs resulting in ToCs close to a No-AD zone also mean more complex and possibly unsafe interactions between MVs and CV/CAVs. Therefore preserving a smooth and safe traffic flow should be a priority in such scenarios for a TMC. Although limitations of the ToC model, such as not capturing evasive and overtaking manoeuvres, may amplify the adverse downward ToC impacts discussed before, we think that the results presented illustrate the feasibility to accomplish both objectives (efficient traffic + automated driving) concurrently as long as demand and traffic mix do not exceed the capacity limit in the scenario. In that regard, the RL method achieves a better overall performance.

Accordingly, this work shows that an AI-based approach is very well able to be on par with more traditional approaches, which usually require a lot of traffic domain knowledge to be developed. Thus, AI-approaches open the promising perspective of control solutions with a limited need for such expertise, and offering a potential for synergies with developments in other areas. It is still a challenging and time-consuming task, though, to develop adequate RL models, which usually require a careful tailoring to the given problem. Besides, the duration of the development cycle of adapting the MDP setup and assessing the learning progress of the altered model is often governed by the computational complexity of the task. In our case, the training of a model variant took about 24 hours (on a Intel Core i9-10900X CPU 10 × 3.7GHz) and numerous iterations were necessary to find an adequate formalization and parametrization of the training experiment.

Finally, we plan to adapt our RL-based models in future work so that they also account for safety objectives, and conduct an explicit microscopic traffic simulation based safety evaluation that will encompass an in-depth analysis of conflicts and relevant surrogate safety assessment measures (SSMs). Moreover, the aspect on how to handle MRMs in such traffic scenarios, which were mostly ignored in this study, might be a future research focus for applying RL to elaborate TM schemes, but it will take significant effort and time, considering the rather low sample efficiency of the training process. In addition, we think that not much more can be gained in terms of traffic efficiency with any other TM approach, for a scenario such as we presented, since a control transition itself, as modelled here, diminishes capacity, and there is no method that can bring it back.

ACKNOWLEDGEMENT

Open access publication fees were covered by the DLR Publication Fund.

REFERENCES

[1] On-Road Automated Driving (ORAD) Committee (SAE Int., Warrendale, PA, USA). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, (Jun. 2018). [Online]. Available: https://doi.org/10.4271/J3016_201806

[2] F. Favaró, S. Eurich, and N. Nader, "Autonomous vehicles disengagements: Trends, triggers, and regulatory limitations," *Accid. Anal. Prevent.*, vol. 110, pp. 136–148, Jan. 2018. [Online]. Available: https://doi.org/10.1016/j.aap.2017.11.001

[3] Z. Lu, R. Happee, C. D. D. Cabrall, M. Kyriakidis, and J. C. F. de Winter, "Human factors of transitions in automated driving: A general framework and literature survey," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 43, pp. 183–196, Nov. 2016. [Online]. Available: https://doi.org/10.1016/j.trf.2016.10.007

[4] V. Melcher, S. Rauh, F. Diederichs, H. Widlroither, and W. Bauer, "Take-over requests for automated driving," *Procedia Manuf.*, vol. 3, pp. 2867–2873, Jan. 2015. [Online]. Available: https://doi.org/10.1016/j.promfg.2015.07.788

[5] S. Petermeijer, P. Bazilinskyy, K. Bengler, and J. de Winter, "Take-over again: Investigating multimodal and directional tors to get the driver back into the loop," *Appl. Ergonom.*, vol. 62, pp. 204–215, Jul. 2017. [Online]. Available: https://doi.org/10.1016/j.apergo.2017.02.023

[6] M. Bahram, M. Aeberhard, and D. Wollherr, "Please take over! an analysis and strategy for a driver take over request during autonomous driving," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2015, pp. 913–919, doi: 10.1109/IVS.2015.7225801.

[7] S. A. M. Agriesti, M. Ponti, G. Marchionni, and P. Gandini, "Cooperative messages to enhance the performance of L3 vehicles approaching roadworks," *Eur. Transp. Res. Rev.*, vol. 13, p. 1, Jan. 2021. [Online]. Available: https://doi.org/10.1186/s12544-020-00457-z

[8] R. Bellman, "A Markovian decision process," *J. Math. Mech.*, vol. 6, no. 5, pp. 679–684, 1957. [Online]. Available: http://www.jstor.org/stable/24900506

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[10] M. Coşkun, A. Baggag, and S. Chawla, "Deep reinforcement learning for traffic light optimization," in *Proc. IEEE Int. Conf. Data Min. Workshops (ICDMW)*, Singapore, 2018, pp. 564–571, doi: 10.1109/ICDMW.2018.00088.

[11] F. Belletti, D. Haziza, G. Gomes, and A. M. Bayen, "Expert level control of ramp metering based on multi-task deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1198–1207, Apr. 2018, doi: 10.1109/TITS.2017.2725912.

[12] Q. Guo, O. Angah, Z. Liu, and X. J. Ban, "Hybrid deep reinforcement learning based eco-driving for low-level connected and automated vehicles along signalized corridors," *Transp. Res. C, Emerg. Technol.*, vol. 124, Mar. 2021, Art. no. 102980. [Online]. Available: https://doi.org/10.1016/j.trc.2021.102980

[13] D. M. Vlachogiannis, E. I. Vlahogianni, and J. Golias, "A reinforcement learning model for personalized driving policies identification," *Int. J. Transp. Sci. Technol.*, vol. 9, no. 4, pp. 299–308, 2020. [Online]. Available: https://doi.org/10.1016/j.ijtst.2020.03.002

[14] D. Nallaperuma *et al.*, "Online incremental machine learning platform for big data-driven smart traffic management," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4679–4690, Dec. 2019, doi: 10.1109/TITS.2019.2924883.

[15] "Easy Tech: Conditionally Automated Driving With the Drive Pilot." Jul. 2021. [Online]. Available: https://www.daimler.com/magazine/technology-innovation/easy-tech-drive-pilot.html (Accessed: Dec. 13, 2021).

[16] F. M. Favaró, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju, "Examining accident reports involving autonomous vehicles in California," *PLoS One*, vol. 12, pp. 1–20, Sep. 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0184952

[17] S. Wang and Z. Li, "Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data," *Accid. Anal. Prevent.*, vol. 129, pp. 44–54, Aug. 2019. [Online]. Available: https://doi.org/10.1016/j.aap.2019.04.015

[18] A. M. Boggs, R. Arvin, and A. J. Khattak, "Exploring the who, what, when, where, and why of automated vehicle disengagements," *Accid. Anal. Prevent.*, vol. 136, Mar. 2020, Art. no. 105406. [Online]. Available: https://doi.org/10.1016/j.aap.2019.105406

[19] Z. H. Khattak, M. D. Fontaine, and B. L. Smith, "Exploratory investigation of disengagements and crashes in autonomous vehicles under mixed traffic: An endogenous switching regime framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7485–7495, Dec. 2021. [Online]. Available: https://doi.org/10.1109/TITS.2020.3003527

[20] A. D. McDonald *et al.*, "Toward computational simulations of behavior during automated driving takeovers: A review of the empirical and modeling literatures," *Human Factors J. Human Factors Ergonom. Soc.*, vol. 61, no. 4, pp. 642–688, 2019. [Online]. Available: https://doi.org/10.1177/0018720819829572

[21] X. Xin *et al.*, "A literature review of the research on take-over situation in autonomous driving," in *Proc. Int. Conf. Human-Comput. Interact.*, 2019, pp. 160–169, doi: 10.1007/978-3-030-23538-3_12.

[22] J. W. Van Lint and S. C. Calvert, "A generic multi-level framework for microscopic traffic simulation—Theory and an example case in modelling driver distraction," *Transp. Res. B, Methodol.*, vol. 117, pp. 63–86, Nov. 2018. [Online]. Available: https://doi.org/10.1016/j.trb.2018.08.009

[23] S. C. Calvert and B. van Arem, "A generic multi-level framework for microscopic traffic simulation with automated vehicles in mixed traffic," *Transp. Res. C, Emerg. Technol.*, vol. 110, pp. 291–311, Jan. 2020. [Online]. Available: https://doi.org/10.1016/j.trc.2019.11.019

[24] L. Lücken, E. Mintsis, K. Porfyri, R. Alms, Y.-P. Flötteröd, and D. Koutras, "From automated to manual—Modeling control transitions with sumo," in *Proc. SUMO User Conf.*, 2019, pp. 124–144. [Online]. Available: https://doi.org/10.29007/sfgk

[25] P. A. Lopez *et al.*, "Microscopic traffic simulation using SUMO," in *Proc. 21st IEEE Int. Conf. Intell. Transp. Syst.*, Nov. 2018, pp. 2575–2582. [Online]. Available: https://elib.dlr.de/127994/

[26] E. P. Todosiev, "The action point model of the driver-vehicle system," Ph.D. dissertation, Dept. Doctor Philos., Ohio State Univ., Athens OH, USA, 1963. [Online]. Available: http://rave.ohiolink.edu/etdc/view?acc_num=osu1486555089597102

[27] W. Xin, J. Hourdos, P. Michalopoulos, and G. Davis, "The less-than-perfect driver: A model of collision-inclusive car-following behavior," *Transp. Res. Rec.*, vol. 2088, no. 1, pp. 126–137, 2008. [Online]. Available: https://doi.org/10.3141/2088-14

[28] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences*, 4th ed. Berlin, Germany: Springer, 2009. [Online]. Available: https://link.springer.com/book/9783540707127

[29] M. Treiber and A. Kesting, *Traffic Flow Dynamics: Data, Models and Simulation*, 1st ed. Heidelberg, Germany: Springer, 2013. [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-32460-4#about

[30] E. Mintsis *et al.* "TransAID Deliverable 3.1—Modelling, Simulation and Assessment of Vehicle Automations and Automated Vehicles' Driver Behaviour in Mixed Traffic." 2019. [Online]. Available: https://cordis.europa.eu/project/id/723390/results

[31] S. Maerivoet *et al.* "TransAID Deliverable 4.2—Preliminary Simulation and Assessment of Enhanced Traffic Management Measures." 2019. [Online]. Available: https://cordis.europa.eu/project/id/723390/results

[32] B. Coll-Perales *et al.*, "Prototyping and evaluation of infrastructure-assisted transition of control for cooperative automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 4, 2021. [Online]. Available: https://doi.org/10.1109/TITS.2021.3061085

[33] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.

[34] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.

[35] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3152–3168, Aug. 2020. [Online]. Available: https://doi.org/10.1109/TITS.2019.2929020

[36] T. Chu, J. Wang, L. Codecá, 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. [Online]. Available: https://doi.org/10.1109/TITS.2019.2901791

[37] J. V. S. Busch, V. Latzko, M. Reisslein, and F. H. P. Fitzek, "Optimised traffic light management through reinforcement learning: Traffic state agnostic agent vs. holistic agent with current V2I traffic state knowledge," *IEEE Open J. Intell. Transp. Syst.*, vol. 1, pp. 201–216, 2020. [Online]. Available: https://doi.org/10.1109/OJITS.2020.3027518

[38] H. Wang, H. Chen, Q. Wu, C. Ma, and Y. Li, "Multi-intersection traffic optimisation: A benchmark dataset and a strong baseline," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 126–136, 2021, doi: 10.1109/OJITS.2021.3126126.

[39] E. Vinitsky *et al.*, "Benchmarks for reinforcement learning in mixed-autonomy traffic," in *Proc. 2nd Conf. Robot Learn.*, Oct. 2018, pp. 399–409. [Online]. Available: http://proceedings.mlr.press/v87/vinitsky18a.html

[40] Z. Li, P. Liu, C. Xu, H. Duan, and W. Wang, "Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3204–3217, Nov. 2017. [Online]. Available: https://doi.org/10.1109/TITS.2017.2687620

[41] G. Bacchiani, D. Molinari, and M. Patander, "Microscopic traffic simulation by cooperative multi-agent deep reinforcement learning," 2019, *arXiv:1903.01365*.

[42] *Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service*, ETSI Standard EN 302 637-2, Apr. 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_en/302600_302699/30263702/01.04.01_60/en_30263702v010401p.pdf

[43] "Intelligent transport systems (ITS); vehicular communications; informative report for the maneuver coordination service, draft 0.0.4," ETSI, Sophia Antipolis, France, ETSI Rep. TR 103 578, 2019.

[44] R. Alms, Y.-P. Flötteröd, E. Mintsis, S. Maerivoet, and A. Correa, "Traffic management for connected and automated vehicles on urban corridors—Distributing take-over requests and assigning safe spots," in *Proc. 3rd Symp. Manag. Future Motorway Urban Traffic Syst.*, Jul. 2020, pp. 1–11. [Online]. Available: https://elib.dlr.de/134136/

[45] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," 2018, *arXiv:1802.09477*.

[46] A. Hill *et al.* "Stable Baselines." 2018. [Online]. Available: https://github.com/hill-a/stable-baselines

[47] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, May 2016. [Online]. Available: http://arxiv.org/abs/1509.02971

[48] A. Noulis, "Reinforcement learning in traffic control for connected automated vehicles," M.S. thesis, Dept. Inst. Transp. Syst., TU Berlin, Berlin, Germany, Oct. 2020. [Online]. Available: https://elib.dlr.de/139169/

[49] V. Milanés and S. E. Shladover, "Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data," *Transp. Res. C, Emerg. Technol.*, vol. 48, pp. 285–300, Nov. 2014. [Online]. Available: https://doi.org/10.1016/j.trc.2014.09.001

[50] E. Vinitsky, N. Lichtle, K. Parvate, and A. Bayen, "Optimizing mixed autonomy traffic flow with decentralized autonomous vehicles and multi-agent RL," 2020, *arXiv:2011.00120*.