Multi-agent reinforcement learning for autonomous vehicles: a survey

Joris Dinneweth, Abderrahmane Boubezoul, René Mandiau, Stéphane Espie

Autonomous Intelligent Systems

REVIEW Open Access

Multi-agent reinforcement learning for autonomous vehicles: a survey

Joris Dinneweth[1,2*], Abderrahmane Boubezoul[1], René Mandiau[3] and Stéphane Espié[1]

Abstract

In the near future, autonomous vehicles (AVs) may cohabit with human drivers in mixed traffic. This cohabitation raises serious challenges, both in terms of traffic flow and individual mobility, as well as from the road safety point of view. Mixed traffic may fail to fulfill expected security requirements due to the heterogeneity and unpredictability of human drivers, and autonomous cars could then monopolize the traffic. Using multi-agent reinforcement learning (MARL) algorithms, researchers have attempted to design autonomous vehicles for both scenarios, and this paper investigates their recent advances. We focus on articles tackling decision-making problems and identify four paradigms. While some authors address mixed traffic problems with or without social-desirable AVs, others tackle the case of fully-autonomous traffic. While the latter case is essentially a communication problem, most authors addressing the mixed traffic admit some limitations. The current human driver models found in the literature are too simplistic since they do not cover the heterogeneity of the drivers' behaviors. As a result, they fail to generalize over the wide range of possible behaviors. For each paper investigated, we analyze how the authors formulated the MARL problem in terms of observation, action, and rewards to match the paradigm they apply.

Keywords: Multi-agent reinforcement learning, Simulation, Autonomous Vehicles

1 Introduction

According to the world health organization (WHO[1]), road accidents kill 1.3 million people and injure 50 million people each year. Several technologies have been proposed to make driving safer, such as advanced driver assistance systems (ADAS), adaptative cruise control (ACC), and intelligent transportation systems (ITS). The latter, with the recent technological advances in communication systems, paved the way for the deployment of autonomous vehicles.

Trommer et al. [1] described five levels of vehicle automation in their technical report, ranging from superficial assistance (level 1) to full automation (level 5). With effective algorithms that prevent fatal accidents, the latter level could make traffic safer. AVs and humans may cohabit in mixed traffic before reaching full automation. However, the evidence suggests that accident-free mixed traffic may be impossible [2]. Human drivers follow informal and subjective norms, but autonomous vehicles comply with traffic rules [3, 4]. Because of their divergent concerns, AVs are unlikely to be effective in mixed traffic. By contrast, coordinating a fully-autonomous fleet is straightforward because AVs act homogeneously and are therefore predictable. AVs should be capable of handling all traffic scenarios, whether they are driving in mixed traffic or fully autonomous fleets. However, because these scenarios are nearly endless, designing ruled-based models is practically certain to fail.

With advances in hardware, machine learning approaches provide new opportunities to generalize driving scenarios. Reinforcement learning (RL) approaches, in particular, are successful at solving sequential decision-making problems, such as Go, Chess, arcade games, and real-time video games [5–9]. In RL, an agent learns and self-corrects by receiving feedback on the quality of its in-

*Correspondence: joris.dinneweth@univ-eiffel.fr
[1]TS2-MOSS, Univ. Gustave Eiffel, 77454, Champs-sur-Marne, France
[2]ENS Paris-Saclay, CNRS, SATIE, Université Paris-Saclay, 91190, Gif-sur-Yvette, France
Full list of author information is available at the end of the article

[1]https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

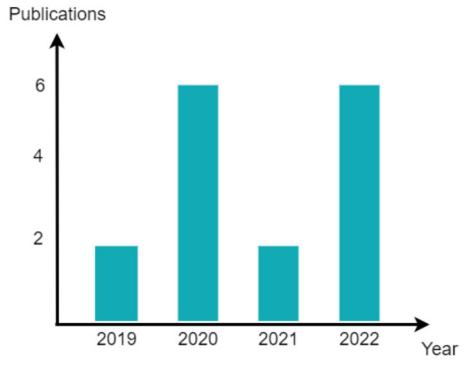


Figure 1 Distribution of the reviewed papers

teractions within an environment. Multi-agent RL (MARL) is a more distributed framework in which several agents simultaneously learn cooperative or competitive behavior. Since several decision-makers learn simultaneously and possibly coordinate, more robust and convincing policies can emerge than with single-agent RL approaches.

Several surveys have investigated relative aspects of RL for AVs more global way. Schmidt et al. [10] tackled autonomous mobility, including traffic management, unmanned aerial vehicles (UAVs), AVs, and resource optimization using MARL algorithms. Elallid et al. [11] surveyed AVs' scene understanding, decision-making, planning, and social behavior using RL approaches. Kiran et al. [12] tackled scene understanding, decision-making, and planning using RL algorithms. Ye et al. [13] tackled motion planning and control using RL approaches. Notwithstanding, no reviews investigated the decision-making of autonomous vehicles using MARL algorithms.

Our survey seeks to fulfill this gap by answering two research questions: (*RQ1*) what is the recent state-of-art of AVs' decision-making using MARL algorithms; and (*RQ2*) what are the topic's primary current limitations. To answer these questions as concisely as possible while considering recent breakthroughs in MARL algorithms, we have restricted this review to sixteen papers published since 2019 (distribution in Fig. 1). We focus our survey on decision-making problems; nonetheless, interested readers can find in [14], a recent survey that focuses on autonomous driving policy learning using deep reinforcement learning (DRL) and deep imitation learning (DIL) techniques.

We have organized the remainder of this review as follows. Firstly, we introduce the state-of-art of RL and MARL algorithms (Sect. 2). Secondly, we highlight the learning schemes and strategies of MARL algorithms (Sect. 3). Thirdly, we review the driving simulation environments (Sect. 4). Fourthly, we investigate articles tackling AVs' decision-making using MARL algorithms (Sect. 5). Lastly, we discuss open challenges and conclude this study (Sect. 6).

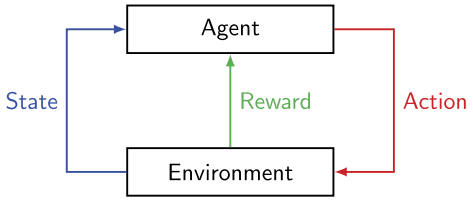


Figure 2 Single-agent reinforcement learning

2 Reinforcement learning

This section provides a state-of-art of single (2.1) and multi-agent (2.2) reinforcement learning algorithms.

2.1 Single-agent reinforcement learning

Reinforcement learning (RL) is a trial-and-error learning method where an agent interacts within an environment [5] (Fig. 2). The agent's goal is to reach the most rewarding states of the environment. The agent explores the environment, grasping its dynamics and devising an appropriate policy (behavior) to discover these states. As a result, the agent gains knowledge from its actions and maximizes long-term accumulated rewards. Non-learning agents who obey stationary policies may be present in the environment. The environment, the state, the actions, and the rewards for an autonomous car may correspond to the roadway, the positions of other vehicles, accelerating or braking, and collision avoidance, respectively.

There are three types of RL learning algorithms: value-based, policy-based, and actor-critic. In value-based methods, the agent implicitly learns a deterministic policy by picking higher-valued actions via a value function that maps state-action pairs. Nevertheless, the value function becomes inefficient as the state-action space grows, such as discrete spaces [15]. In policy-based methods, the agent explicitly learns a stochastic policy function. However, policy-based approaches suffer from high variance, which slows down the learning process. Actor-critic approaches appear to be a reasonable compromise that combines the benefits of the preceding methods. The latter is divided into a critic part which approximates the value function, while an actor part learns a policy based on critic estimations to alleviate the variance. Because they work effectively in real-world contexts with continuous space, actor-critic approaches are widespread within the RL community.

We briefly describe the single-agent RL algorithms (Fig. 3) addressed in Sect. 5. Deep Q-network (DQN) [16] is a value-based agent that builds a deep learning model to estimate future rewards and execute behaviors that lead to the best outcome. Advantage actor-critic (A2C) [17] is an actor-critic agent that builds a stochastic policy to estimate the advantage of taking action over others. Deep deterministic policy gradient (DDPG) [18] is an A2C agent with de-

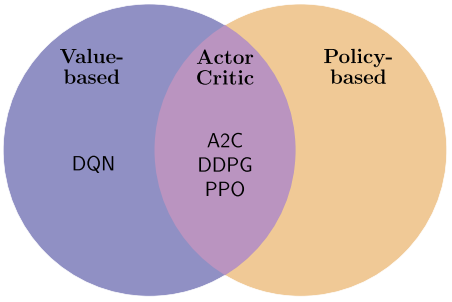


Figure 3 Taxonomy of RL methods

terministic off-policy, which means that the present policy does not guide the learning process. Instead of employing a logarithmic update, proximal policy optimization (PPO) [19] is an expansion of the A2C agent that updates the policy based on the ratio between the old and new policies weighted by the advantage. None of them deal with policy-based methods.

2.2 Multi-agent reinforcement learning

Multi-agent reinforcement learning (MARL) algorithms involve several agents learning simultaneously in a shared environment. Agents are either cooperative, competitive, or have a mixed approach. Cooperative agents possibly communicate to coordinate their actions (Fig. 4) and often share a common reward function. Conversely, competitive agents play a zero-sum game attempting to outperform their opponents. When agents do not behave fully cooperatively or fully competitively, they follow the mix setting, a general-sum game without any restrictions on agents' relations [20].

MARL algorithms follow the same taxonomy of single-RL methods introduced in Fig. 3. Multi-agent extensions of single-agent algorithms are often prefixed with *MA*, e.g., MAA2C and MADDPG [21, 22]. MARL algorithms are more complicated than single-agent RL approaches because several agents learn simultaneously and constantly co-adapt their policies. This non-stationarity disrupts the dynamics of the environment and impedes the learning process [23]. Furthermore, as the number of agents increases, the space expands exponentially, slowing the learning process. The latter phenomenon is called the curse of dimensionality.

In other environments, agents operate with just partial observations of the present state, making learning more challenging; for example, it is hard to observe the whole traffic flow in road driving. To dispel these obstruction zones, agents can communicate in cooperative tasks [24]. Connected autonomous vehicles, for example, could share and merge their local observations to better represent traffic, potentially revealing a vehicle in a blind spot. Non-stationarity and partial observability are mitigated by communication.

Many learning schemes and strategies have been proposed in response to the additional challenges of MARL algorithms, which are exacerbated by the number of agents.

3 Learning schemes

The curse of dimensionality, partial observability, and non-stationarity represent three critical challenges for MARL development. This section introduces how MARL centralized or decentralized the learning and its execution (3.1) and what are learning schemes (3.2) implemented in the reviewed papers that tackle these challenges.

3.1 Centralization and decentralization

In learning algorithms, an agent learns a policy during a training phase and follows it during the execution phase. These phases, in MARL algorithms, can be either centralized or decentralized. In the centralized one, agents share information to improve their policies, whereas, in the decentralized one, they learn independently with no additional information. Three major learning schemes have been proposed depending on whether the training and execution phases are centralized or decentralized.

3.1.1 Centralized training centralized execution (CTCE)

In centralized training centralized execution (CTCE) scheme, a central learner gathers information from agents to learn a joint policy, which mitigates the partial observability and non-stationarity issues. However, CTCE suffers from centralization, which exacerbates the curse of dimensionality. Furthermore, agents with competing goals may disrupt each other's policies, making learning harder. Single-agent RL algorithms may suffice because CTCE does not expressly assume decentralization. In contrast to CTCE, a fully-decentralized scheme has been proposed.

3.1.2 Decentralized training decentralized execution (DTDE)

Decentralized training decentralized execution (DTDE) scheme allows each agent to learn independently without exchanging additional information. As a result, agents are unaware of one another's existence, and the environment appears non-stationary from their viewpoints. Furthermore, Gupta et al. [25] demonstrated that DTDE scales poorly with agent number.

One last scheme has been proposed as an intermediary solution, given the previous limitations of the fully-centralized and fully-decentralized approaches.

3.1.3 Centralized training decentralized execution (CTDE)

Lowe et al. [22] introduced the centralized training decentralized execution (CTDE) method, which overcomes the

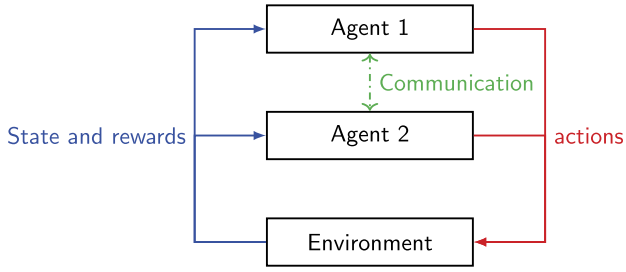
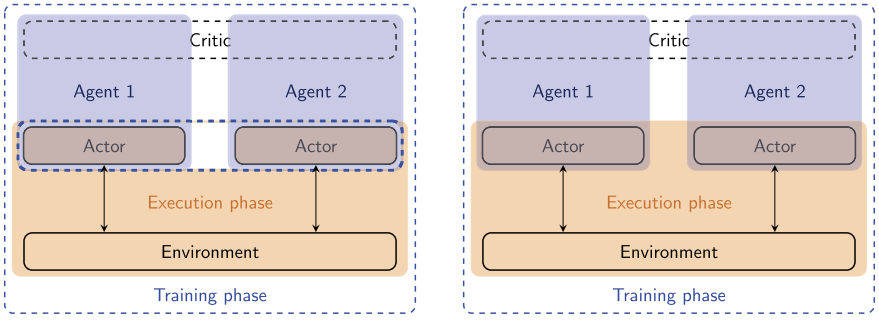


Figure 4 MARL with two communicative agents



(a) Parameter sharing (b) Centralized actor decentralized critic

Figure 5 CTDE learning schemes

shortcomings of the fully-centralized and fully-decentralized approaches. During the training phase, agents share additional information to reduce non-stationarity and partial observability, then discard it during the execution phase. CTDE scheme includes two popular strategies that can be used depending on the agents' nature [25].

Parameter sharing Parameter sharing (PS) is a well-known approach for dealing with large-scale environments where several homogeneous agents cooperate [25]. PS mitigates the curse of dimensionality by allowing all agents to learn simultaneously using a single neural network during the training phase (Fig. 5(a)).

Centralized critic decentralized actor However, when agents are heterogeneous, the centralized critic decentralized actor is more convenient [22]. It follows the actor-critic architecture. Since the critic focuses on assessing the actor, it is no longer helpful for the execution phase. Therefore, each agent receives a duplicate of the actor after the training phase (Fig. 5(b)).

MARL research is still in its infancy, and we have barely skimmed its surface. Interested readers may find comprehensive reviews dedicated to MARL algorithms and challenges [20, 26–30]. In addition to these MARL learning schemes, various RL strategies may overcome multi-agent challenges.

3.2 Learning strategies

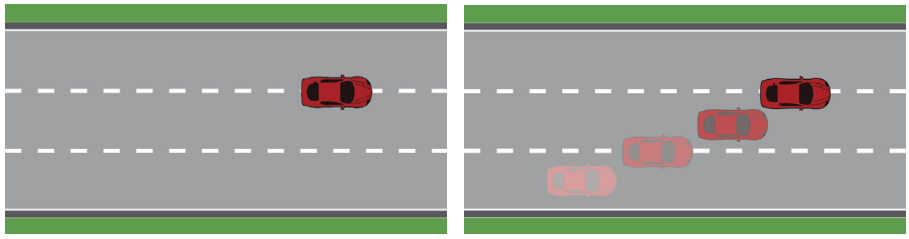
This subsection presents some RL strategies inspired by human cognitive mechanisms that were used in the papers discussed in Sect. 5.

3.2.1 Memory

Memory is a mechanism allowing humans to analyze dynamics. Because RL approaches deal with sequential problems, giving agents memory strengthens their ability to figure out the environment's dynamics [31]. Researchers designed a Recurrent Neural Network (RNN), a memory-based neural network with information cycles that remember the past inputs and reuse them in subsequent decisions. As a result, RNN reduces non-stationarity by improving the analysis of current dynamics based on these experiences. In the case of driving, the memory enables determining the heading of a vehicle between two lanes (Fig. 6).

3.2.2 Masking

Masking prevents humans from performing undesirable actions, making the environment safer and decision-



(a) Without memory (b) With memory

Figure 6 The benefit of memory. It is impossible to figure out the car's heading without memory (6(a)), while it becomes straightforward with memory (6(b))

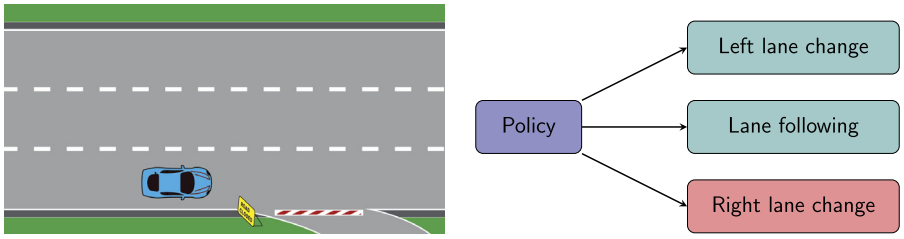


Figure 7 Masking prevents from undertaking undesirable actions (red)

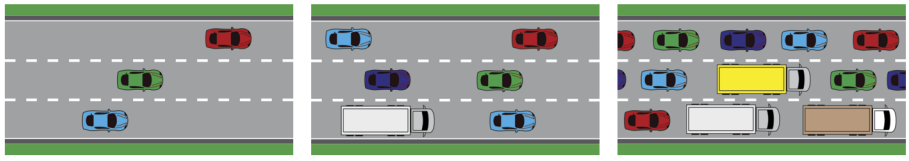


Figure 8 Curriculum Learning for driving. From left to right, agents start in light traffic, increase the complexity, and become more robust in dense traffic

making straightforward [31]. When a designer knows *a priori* that an action is counterproductive, he or she can prevent the agents from undertaking it. For example, when a road is under construction, barriers prevent us from taking it (Fig. 7). Masking speeds up the training and alleviates the curse of dimensionality by narrowing the action space. Another way to ease learning is to reduce exploration.

3.2.3 *Curriculum learning*

Curriculum learning [32] refers to a learning method that gradually increases the difficulty. For example, when people learn to drive, they usually start in low-traffic areas, and when they master it, they move on to denser areas (Fig. 8). In MARL, agents often fail to learn practical policies because of the non-stationarity. With curriculum learning, agents start learning in stationary environments and gradually remove this stationarity, making the task more challenging. Another way to ease learning is to consider it hierarchically.

3.2.4 *Hierarchical reinforcement learning* (*HRL*)

Hierarchical reinforcement learning are "divide and conquer" algorithms [33]. Dividing the main policy into lower-level sub-policies make problems more manageable since these sub-policies can be reused in related tasks (Fig. 9). For example, a left lane change on a highway can reuse the knowledge acquired from a similar task on a country road. Sub-tasks are sometimes less resource-intensive than global tasks; because they can operate in a narrowed state-action space, thus alleviating the curse of dimensionality.

We showed in this section that centralized and decentralized schemes suffer from many problems that learning strategies can alleviate. The following section will describe the MARL-based driving simulation environments.

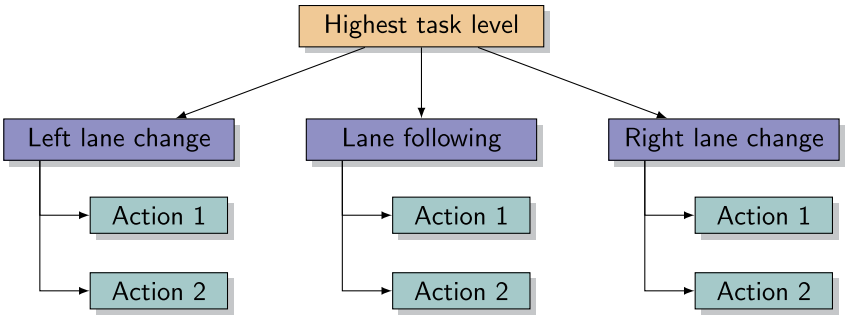


Figure 9 Hierarchical reinforcement learning (inspired from Chen et al. [34]). A higher-level policy (orange) selects a subtask (blue) to perform a sub-policy on a narrow action space (green)

4 MARL-based driving simulation environments

Coordinating a fully-autonomous fleet, i.e., without human drivers, is more tractable than driving in mixed traffic because of the predictable nature of homogeneous agents. Furthermore, to keep traffic flowing, AVs share information and coordinate within short reaction times. Most MARL training use simulation environments (4.1) to learn these features on various scenarios (4.2) and with human driver models (4.3).

4.1 Simulation environments

Simulation environments provide tools to simulate traffic and develop learning algorithms for AVs. They allow benchmarking of the effectiveness of the suggested algorithms before shifting to a real-world implementation. We briefly introduce, in alphabetic order, four simulation environments used in the papers introduced in Sect. 5.

- CARLA [35] is an open-source road environment based on Unreal Engine.[2] It provides assets to model the road environment and implement perception, planning, and control modules.
- Flow[3] [36] is a framework combining the SUMO traffic simulator [37] and a deep RL library Rllab [38]. It provides many traffic scenarios and supports training involving a fixed number of vehicles.
- Highway-env[4] is an open-source Gym-based platform. It provides road scenarios designed to train AVs' decision-making in mixed traffic. According to Schmidt et al. [10], its performance decreases with the number of vehicles.
- MACAD-Gym[5] [39] is a Gym-based training environment based on CARLA. As its name implies, multi-agent connected autonomous driving (MACAD) allows the implementation of communicative agents.

All these simulation environments support the design of different scenario types.

4.2 Driving scenarios

Most papers focus on narrow scenarios instead of considering overall traffic. We present the traffic scenarios according to their complexity.

1. *Highway driving*. This scenario is commonly accepted as the most straightforward scenario and considers two maneuvers: car-following and lane changing. Mastering these maneuvers, which account for 98% of driver actions, is crucial for safe driving. Robust AVs mastering highway driving should avoid collisions and frequent lane changes, which will affect traffic flow.
2. *Merging and exiting*. These maneuvers are similar to lane change but are constrained in space and time. Robust AVs must anticipate gaps in traffic to merge smoothly within the traffic flow and space-time constraints. Inference capabilities should also determine whether a driver is inclined to engage in an altruistic behavior by leaving a gap, which is not straightforward since AVs are agnostic about informal rules.
3. *Intersections and roundabouts*. There are heterogeneous configurations of intersection, and apprehending them can be challenging. For instance, designers failed to generalize them via rules-based models and designed a decision graph for each one, which is tedious. Robust AVs should generalize them and figure out the singularities of each.

Because designing a generic model of intersections is difficult, most research concentrate on the first two levels. Regardless of the scenario, robust AVs should have an advantage by avoiding more collisions if they can predict human driver behavior.

[2] www.unrealengine.com

[3] https://github.com/flow-project

[4] https://github.com/eleurent/highway-env

[5] https://github.com/praveen-palanisamy/macad-gym

4.3 Human driver models

AVs have difficulty adapting to the heterogeneity of human behavior because it produces additional uncertainty and forces caution. To overcome uncertainty, humans make assumptions based on experience, informal rules, and behavioral cues, which are sometimes biased or stereotyped [40, 41]. It is impossible to replicate the entire human cognitive process, and therefore, AVs often learn with oversimplified human models.

Human-driven vehicle (HDV) models simulate car-following and lane-changing maneuvers [42–44]. The well-known intelligent driver model (IDM) describes speed and acceleration based on the driver's preferences for speed and headway [45]. The IDM is often combined with the MOBIL or LC2013 lane-changing model, which considers the utility and risk associated with this maneuver [46, 47]. Although the literature refers to them as human-driven models, they lack human traits such as psychology or intrinsic motivation.

Designing AVs for mixed traffic is challenging because of the fundamental differences between humans and machines. Although inferring human social behavior helps AVs' decision-making, the following section shows that this approach is not widespread in the literature.

5 MARL algorithms for AVs

We have identified four research paradigms throughout the MARL decision-making for AVs literature. Some authors focused on mixed traffic where AVs drive in a self-concern way (5.1), while others attempted to incorporate social abilities into their decision-making (5.2). In both cases, the authors realized that the current HDV models do not fulfill their objectives since they are oversimplified and fail at providing a heterogeneity of behaviors. As a result, researchers designed a more sophisticated HDV model endowed with social capabilities (5.3). The last paradigm tackles the fully autonomous traffic case where no human driver can disturb AVs' coordination (5.4). Finally, we present the formulation of the authors (5.5). For each paradigm introduced, we present the authors' formulation of the MARL problem in terms of observation, action, and reward function.

5.1 Mixed traffic

Before reaching the full automation level, AVs will potentially cohabit with human drivers in mixed traffic, which is no easy feat. AVs follow homogeneous policies, while humans are sometimes erratic and irrational. Here, we focus on papers suggesting self-concern AVs driving in mixed traffic.

Wang et al. [48] trained AVs on three scenarios: a ring network, a figure-of-eight network, and a mini-city with intersections and roundabouts. The ego-agent state comprises its position, speed, and the distance and speed headway of the leading and following vehicles. AVs communicate local observations with other AVs within range. The ego-agent's actions are constrained within predefined discrete acceleration values, and its reward function promotes safety and efficiency.

Dong et al. [31] tackled a challenging environment where AVs have to exit by one of the two off-ramps on a three-lane highway. The agent's observations contain the relative speeds, longitudinal locations, lane positions, and intentions of surrounding AVs, as well as an adjacency matrix and a mask. AVs pick up high-level actions: lane change or lane keeping. Functions reward when each AV reaches the desired off-ramp indicated by the intention and penalizes collision and lane changes to prevent versatility.

Han and Wang [49] trained AVs to drive on a three-lane freeway. Each AV observes its position, velocity, acceleration, and data captured from an onboard camera and LIDAR sensors. Additionally, AVs share their states, actions, and observations with each other. AVs select high-level actions such as lane keeping, lane change, or emergency stop and are rewarded according to their velocities and passengers' comfort. The reward system deals with the credit assignment problem, i.e., how to fairly reallocate a shared global reward by marginalizing rewards using the Shapley value. In the cooperative game theory, the Shapley value is a solution concept that distributes fair payoffs to players proportionally to their contribution. Since the complexity of the Shapley value is polynomial with the number of agents, the authors estimated via a neural network and extended it to sequential problems.

AVs decision-making in mixed traffic is significantly impacted by the absence of other AVs in their vicinities. As AVs communicate local observations within range, meaning the uncertainty about the environment grows as the number of surrounding AVs decreases. To overcome this challenge, some researchers envision AVs that are more aware of their surroundings and propose algorithms with social capabilities.

5.2 Socially desirable AVs

Socially desirable AVs will likely include the concept of altruism. In psychology, social value orientation (SVO) quantifies an individual's level of altruism, i.e., how much importance to place on others. Lower SVO levels denote selfish behavior, while higher levels denote true altruism.

In their first paper, Toghi et al. [50] tackled the merging and exiting scenarios with socially desirable AVs. AVs observe the kinematics of their neighboring vehicles as well as their last high-level actions to extract the temporal information giving their current trajectories. They perform meta-actions, including lane change, acceleration, and deceleration. The socially desirable behavior is induced through a reward function acting as a trade-off between egoistic and altruistic behavior, differentiating altruism towards AVs and human-driven vehicles. The SVO of

the AV weighs this trade-off and the distance of the surrounding vehicles considered.

In their second paper, Toghi et al. [51] enhanced their approach using a 3D convolution network with the relative vehicle speeds as channels. Further experiments have identified an optimal level of SVO that improves overall traffic flow and show that overly altruistic AVs reduce performance. Their third paper achieves better results using a multi-agent actor-critic algorithm [52].

Chen et al. [53] trained AVs to avoid collisions in a merging scenario using a supervisor prioritizing vehicles that merge because their situation is time-critical. AVs observe lateral and longitudinal positions and velocities of surrounding vehicles and pick up meta-action among lane change, accelerating or decelerating. A reward function promotes fast merging, high velocity, and safe time headway and penalizes collisions. This function is a global reward shared by all the AVs in the simulation for encouraging coordination among AVs.

As the results of these articles noticed, socially desirable AVs improve the success rate of merging and exiting maneuvers. Nonetheless, this coordination is facilitated because human-driven vehicles are all controlled by the IDM model and thus are easily predictable. Designing robust AVs that cope with heterogeneous driver behavior and traffic simulations will require comprehensive driver models.

5.3 Heterogeneous HDVs

Robust AVs inevitably will have to be trained to drive in complex mixed traffic composed of heterogeneous human-driven vehicles (HDV). Some researchers [54] attempted to learn an HDV model via inverse RL (IRL), a technique for figuring out an agent's reward function given its policy; but this approach is highly dependent on the quality of the extracted data and the studied scenario. As a consequence, there is a need for a "realistic" and heterogeneous HDV model.

Valiente et al. [55] extended the research of Toghi et al. by incorporating an SVO factor into the IDM model used for controlled HDVs. Similarly, Zhou et al. [56] endowed HDVs with a politeness factor, and Hu et al. [57] designed a social HDV model with different levels of cooperation.

All the mentioned authors took advantage of their new HDV models by enabling AVs to infer this SVO and thus anticipated which driver is prone to act altruistically or not.

5.4 Fully-autonomous fleet

When AVs reach the fifth level of automation, human drivers might be considered the main threat to road safety and therefore be banned from driving. In this context, all traffic will be composed of fully-autonomous fleets.

Yu et al. [58] addressed the problem of coordination on the highway. AVs observe their current lane position, speed, and the distances and velocities of four neighboring vehicles. Actions comprise driving in the driving lane at a suboptimal speed or driving in the overtaking lane with a higher velocity. The reward function exclusively promotes safety and is shared among a local group of AVs depicted by coordination graphs.

Bhalla et al. [59] learned AVs to better communicate and coordinate on a highway. They measure them against DIAL, a benchmark algorithm that focuses on learning to communicate in cooperative tasks [60]. Unlike DIAL, their method does not require past experiences, which mitigates non-stationarity and stabilizes learning. AVs' actions include sending messages, accelerating, decelerating, and direction change. The reward function does not provide explicit rewards for cooperation between the agents but promotes safety distance and penalizes crashes.

Liu et al. [15] proposed a framework for fleet control where each vehicle learns to maintain a constant headway with the vehicles ahead and behind on a highway. Each AV observes its position and speed, as well as those of front and rear vehicles. To maintain the homogeneity of the fleet, a reward function penalizes the AVs which are not at equidistance to the front and rear vehicles or AVs whose velocity and acceleration differ from the group.

Palanisamy [39] designed MACAD, a simulation environment to simulate AV's perception, decision-making, and control. In an intersection scenario, AVs' observations are images captured from an onboard camera, and they can pick up one of the eight discrete actions controlling steering angle, throttle, and brake. The function rewards AVs crossing the intersection while maintaining a high speed and avoiding collisions. Optionally a factor encourages/discourages cooperativeness/competitiveness among the agents.

Nakka et al. [61] tackled the coordination problem in a merging scenario. The merging AV observes the distances and velocities of the surrounding vehicles and the distance from the end of the merging zone. Actions allow the AV to accelerate or decelerate, and the reward function encourages agents to maintain their speed within a predefined range and penalizes rear-end collisions.

5.5 Synthesis

We synthesize the previous papers according to the concepts introduced in this survey (Table 1). Most authors used single-agent RL methods, especially those based on DQN, to address MARL problems (12 out of 16) and mainly adopted the CTDE scheme for MARL approaches (3 out of 4). The action space's nature seems to guide the motivations for using value-based or actor-critic methods since the latter better deal with continuous action space. In addition, few articles used learning strategies or explicitly mentioned them.

Interestingly, most papers (12 out of 16) focused their study on simulations involving few agents (≤ 10). This

Table 1 Summary of papers according to the problem addressed and simulation settings. Scenarios include merging (M), exiting (E), highway (H) without merging nor exiting, urban navigation comprising intersections and roundabouts (U), and intersection (I). Learning strategies include Hierarchical Reinforcement Learning (HRL), Curriculum Learning (CL), Memory module (Mem), and Masking (Mask)

Article	Class	Algorithm	Scheme	No. AVs	HDV model	Scenario	Simulator	Learning strat.
Yu et al. [58]	Fleet	DQN	DTDE	≤ 20	–	H	–	–
Bhalla et al. [59]	Fleet	DQN	CTDE / DTDE	≤ 10	–	H	Gym-based	Mem, HRL
Liu et al. [15]	Fleet	DQN	–	≤ 10	–	H	–	–
Palanisamy [39]	Fleet	IMPALA[1]	CTDE	≤ 5	–	I	MACAD	–
Nakka et al. [61]	Fleet	DDPG	CTDE	≤ 10	–	M	–	–
Wang et al. [48]	Mixed	PPO	–	≤ 10	IDM	U	Flow	–
Dong et al. [31]	Mixed	DQN	CTDE	≤ 20	IDM-LC2013	E	Flow	Mem, Mask
Han and Wang [49]	Mixed	MADDPG	CTDE	≤ 30	CARLA-autopilot	H	CARLA	–
Toghi et al. [50]	Social	DQN	DTDE	≤ 5	IDM-MOBIL	M	Highway-env	–
Toghi et al. [51]	Social	DQN	DTDE	≤ 5	IDM-MOBIL	M	Highway-env	–
Toghi et al. [52]	Social	MA2C	DTDE	≤ 5	IDM-MOBIL	M / E	Highway-env	–
Chen et al. [53]	Social	MA2C	CTDE	≤ 5	IDM-MOBIL	M	Highway-env	CL, Mask
Valiente et al. [55]	HDV	DQN	DTDE	≤ 5	IDM-MOBIL[2]	M / E	Highway-env	–
Zhou et al. [56]	HDV	A2C	DTDE	≤ 5	IDM-MOBIL[2]	H	Highway-env	–
Hu et al. [57]	HDV	MA2C	CTDE	≤ 10	IDM[2]	M	–	CL

[1] Single-agent actor-critic algorithm designed for multi-task RL [63].
[2] Modified version.

choice is presumably motivated by the MARL challenges, notably the curse of dimensionality [62].

Most studies investigated highway driving and merging scenarios (13 out of 16), as these critical maneuvers involve anticipation and often cause accidents to AVs. For their simulations, Gym-based environments prevail due to their manageable API for RL. Similarly, IDM prevails because of its efficiency and computational simplicity.

Since 2019, few papers have addressed AVs' decision-making using MARL compared to those using single-agent RL. Due to the limited number of articles dealing with MARL, our conclusions may be biased, so we invite readers to consider this.

6 Open challenges and conclusion

Overall, most studies focus on simulations rather than addressing transferability to real traffic scenarios. The needs for "realistic" driver models, safe and interpretable models are two significant problems for AV simulation discussed in this section.

Safety is undoubtedly the critical point of the development of AV algorithms. In MARL, designing a safe policy is a real challenge that implies considering safety constraints at the agent and group levels. The constrained markov decision process (CMDP) framework provides tools for designing such safe RL [64] algorithms.

Most studies agree that existing HDV models are unrealistic because they disregard human characteristics such as psychological and biological traits. Although some researchers tried to provide heterogeneity in HDV models, their models are still limited to a single SVO trait. Besides, despite their differences, HDV and AV models behave deterministically. Introducing AVs trained with these HDV models into real-world traffic would likely result in accidents.

Therefore, developing convincing driver models for safe driving is critical, as driving styles vary among countries and cultures [65]. Attempts have been made using inverse reinforcement learning (IRL), but these algorithms are overly dependent on the situations under study and frequently fail to generalize. Others have proposed utilizing MARL algorithms to learn social norms, which may be a new field of research [66].

Another way to prepare AVs for real-world traffic is to make them trustworthy by incorporating interpretability. Explainable artificial intelligence (EAI) is an important research topic gaining interest over the years, mainly because lawmakers require AI to be interpretable, as in Europe with the general data protection regulation (GDPR[6]). Therefore, robust AVs should incorporate interpretable algorithms providing security and robustness guarantees. Interpreting MARL policies involves explaining short- and long-term decision-making and interactions of multiple agents. This may be accomplished via Causal MARL [67].

Since multi-agent simulations, and MARL algorithms in a broader way, enable the emergence of organizational structures, it might be interesting to investigate how self-organization occurs in a fully autonomous fleet with no predetermined rules. While researchers tend to incorporate standards into AVs' decision-making, they do not rule them out for the fully-autonomous fleets. These emergent organizations may be more appropriate for AVs than current regulations based on humans' limitations.

[6] https://gdpr-info.eu/

We posed two research questions in the introduction (1), which we now address.

- *RQ1*. Recent AVs' decision-making research focused on two paradigms. On the one hand, since autonomous vehicles may soon coexist with human drivers, mixed traffic received much attention. Some studies concentrated on improving traffic safety and throughput, while others proposed empowering AVs with social abilities. Some attempted to design HDV models that mimic driver altruism to robustify AVs' policies. On the other hand, since human drivers might be banned from traffic, some researchers devised fully-autonomous fleets that should enhance the overall traffic flow and security.
- *RQ2*. Designing traffic simulations with adequate HDV models is challenging, and despite the proposed models, none covered the heterogeneity of human behavior. Given the current limitations, it seems involved to consider mixed traffic, and future research will likely pay more attention to this problem. In addition, since intersections and roundabouts are manifolds, most studies concentrated on the most straightforward scenarios, such as highway driving, merging, and exiting. Finally, most experiments involved few agents due to the aforementioned MARL challenges [62].

In conclusion, RL and MARL algorithms have recently received interest due to their recent achievements and generalization capabilities. They provide a practical approach for learning complex policies involving real-time decision-making in stochastic environments. However, many challenges remain in mitigating the scalability when involving numerous agents. Furthermore, mixed traffic does not meet the security standards in the current simulations. Recent papers attempted to mimic human behavior, particularly social capabilities, to enforce AVs' policies. Given current AVs' algorithms, future research will most likely continue to design less deterministic driver models.

Funding
This manuscript was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 815001 (project DriveToTheFuture).

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contribution
Original draft preparation, JD; writing—review and editing, JD, AB, SE and RM. All authors have read and agreed to the published version of the manuscript.

Author details
[1]TS2-MOSS, Univ. Gustave Eiffel, 77454, Champs-sur-Marne, France. [2]ENS Paris-Saclay, CNRS, SATIE, Université Paris-Saclay, 91190, Gif-sur-Yvette, France. [3]CNRS, UMR 8201—LAMIH, Univ. Polytechnique Hauts-de-France, F-59313, Valenciennes, France.

Publisher's Note

Received: 31 May 2022 Revised: 26 October 2022
Accepted: 30 October 2022 Published online: 16 November 2022

References

1. S. Trommer, V. Kolarova, E. Fraedrich, L. Kröger, B. Kickhöfer, T. Kuhnimhof, B. Lenz, P. Phleps, The Impact of Vehicle Automation on Mobility Behaviour. Auton. Driv. 94, (2016)
2. D. Petrović, R. Mijailović, D. Pešić, Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers. Transp. Res. Proc. **45**, 161–168 (2020). https://doi.org/10.1016/j.trpro.2020.03.003
3. G.J. Wilde, Social interaction patterns in driver behavior: an introductory review. Hum. Factors **18**(5), 477–492 (1976)
4. M. Haglund, L. Åberg, Speed choice in relation to speed limit and influences from other drivers. Transp. Res., Part F Traffic Psychol. Behav. **3**(1), 39–51 (2000)
5. R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning Series, 2nd edn. (MIT Press, Cambridge, 2018)
6. D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
7. D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel et al., Mastering chess and shogi by self-play with a general reinforcement learning algorithm (2017). arXiv preprint. arXiv:1712.01815
8. J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, D. Silver, Mastering atari, go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (2020). https://doi.org/10.1038/s41586-020-03051-4
9. O. Vinyals, I. Babuschkin, W.M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev et al., Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)
10. L.M. Schmidt, J. Brosig, A. Plinge, B.M. Eskofier, C. Mutschler, An introduction to multi-agent reinforcement learning and review of its application to autonomous mobility (2022). arXiv preprint. arXiv:2203.07676
11. B.B. Elallid, N. Benamar, A.S. Hafid, T. Rachidi, N. Mrani, A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving. J. King Saud Univ, Comput. Inf. Sci. (2022). https://doi.org/10.1016/j.jksuci.2022.03.013
12. B.R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A.A. Al Sallab, S. Yogamani, P. Pérez, Deep reinforcement learning for autonomous driving: a survey. IEEE Trans. Intell. Transp. Syst. (2021). https://doi.org/10.1109/TITS.2021.3054625
13. F. Ye, S. Zhang, P. Wang, C.-Y. Chan, A survey of deep reinforcement learning algorithms for motion planning and control of autonomous vehicles, in *2021 IEEE Intelligent Vehicles Symposium (IV)* (IEEE Press, New York, 2021), pp. 1073–1080
14. Z. Zhu, H. Zhao, A survey of deep rl and il for autonomous driving policy learning. IEEE Trans. Intell. Transp. Syst. (2021). https://doi.org/10.1109/TITS.2021.3134702
15. B. Liu, Z. Ding, C. Lv, Platoon control of connected autonomous vehicles: a distributed reinforcement learning method by consensus. IFAC-PapersOnLine **53**(2), 15241–15246 (2020)
16. C.J. Watkins, P. Dayan, Q-learning. Mach. Learn. **8**(3), 279–292 (1992)
17. V. Mnih, A.P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in *International Conference on Machine Learning* (PMLR, 2016), pp. 1928–1937

18. T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning (2015). arXiv preprint. arXiv:1509.02971
19. J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz, Trust region policy optimization, in *International Conference on Machine Learning* (PMLR, 2015), pp. 1889–1897
20. K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: a selective overview of theories and algorithms. Handb. Reinf. Learn. Control, 321–384 (2021)
21. T. Chu, J. Wang, L. Codecà, Z. Li, Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans. Intell. Transp. Syst. **21**(3), 1086–1095 (2019)
22. R. Lowe, Y.I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments. Adv. Neural Inf. Process. Syst. 30, (2017). https://doi.org/10.5555/3295222.3295385
23. P. Hernandez-Leal, M. Kaisers, T. Baarslag, E.M. de Cote, A Survey of Learning in Multiagent Environments: Dealing with Non-Stationarity (2019). arXiv:1707.09183 [cs]
24. Y. Shoham, K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations* (Cambridge University Press, USA, 2008)
25. J.K. Gupta, M. Egorov, M. Kochenderfer, Cooperative multi-agent control using deep reinforcement learning, in *International Conference on Autonomous Agents and Multiagent Systems* (Springer, Berlin, 2017), pp. 66–83
26. P. Hernandez-Leal, B. Kartal, M.E. Taylor, A survey and critique of multiagent deep reinforcement learning. Auton. Agents Multi-Agent Syst. **33**(6), 750–797 (2019)
27. T.T. Nguyen, N.D. Nguyen, S. Nahavandi, Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. IEEE Trans. Cybern. **50**(9), 3826–3839 (2020). https://doi.org/10.1109/TCYB.2020.2977374
28. L. Canese, G.C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Re, S. Spanò, Multi-agent reinforcement learning: a review of challenges and applications. Appl. Sci. **11**(11), 4948 (2021). https://doi.org/10.3390/app11114948
29. S. Gronauer, K. Diepold, Multi-agent deep reinforcement learning: a survey. Artif. Intell. Rev. **55**(2), 895–943 (2022). https://doi.org/10.1007/s10462-021-09996-w
30. A. OroojlooyJadid, D. Hajinezhad, A Review of Cooperative Multi-Agent Deep Reinforcement Learning (2021) arXiv:1908.03963 [cs, math, stat]
31. J. Dong, S. Chen, P.Y.J. Ha, Y. Li, S. Labi, A drl-based multiagent cooperative control framework for cav networks: a graphic convolution q network (2020). arXiv preprint. arXiv:2010.05437
32. Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in *Proceedings of the 26th Annual International Conference on Machine Learning—ICML'09* (ACM Press, Montreal, 2009), pp. 1–8. https://doi.org/10.1145/1553374.1553380
33. S. Pateria, B. Subagdja, A.-H. Tan, C. Quek, Hierarchical reinforcement learning: a comprehensive survey. ACM Comput. Surv. (CSUR) **54**(5), 1–35 (2021)
34. Y. Chen, C. Dong, P. Palanisamy, P. Mudalige, K. Muelling, J.M. Dolan, Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019), pp. 1326–1334. https://doi.org/10.1109/CVPRW.2019.00172
35. A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, V. Koltun, Carla: an open urban driving simulator, in *Conference on Robot Learning* (PMLR, 2017), pp. 1–16
36. C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, A.M. Bayen, Flow: architecture and benchmarking for reinforcement learning in traffic control (2017). arXiv preprint. arXiv:1710.05465
37. M. Behrisch, L. Bieker, J. Erdmann, D. Krajzewicz, Sumo–simulation of urban mobility: an overview, in *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation* (ThinkMind, 2011)
38. Y. Duan, X. Chen, R. Houthooft, J. Schulman, P. Abbeel, Benchmarking deep reinforcement learning for continuous control, in *International Conference on Machine Learning* (PMLR, 2016), pp. 1329–1338
39. P. Palanisamy, Multi-agent connected autonomous driving using deep reinforcement learning, in *2020 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Glasgow, 2020), pp. 1–7. https://doi.org/10.1109/IJCNN48605.2020.9207663
40. C. Munduteguy, Reconnaissance d'intention et prédiction d'action pour la gestion des interactions en environnement dynamique. PhD thesis, Paris, CNAM (2001)
41. C. Munduteguy, F. Darses, Perception et anticipation du comportement d'autrui en situation simulée de conduite automobile. Le Trav. Hum. **70**(1), 1–32 (2007)
42. Q. Chao, H. Bi, W. Li, T. Mao, Z. Wang, M.C. Lin, Z. Deng, A survey on visual traffic simulation: models, evaluations, and applications in autonomous driving, in *Computer Graphics Forum*, vol. 39 (Wiley, New York, 2020), pp. 287–308
43. S.P. Hoogendoorn, P.H. Bovy, State-of-the-art of vehicular traffic flow modelling. Proc. Inst. Mech. Eng., Part I, J. Syst. Control Eng. **215**(4), 283–303 (2001)
44. S. Moridpour, M. Sarvi, G. Rose, Lane changing models: a critical review. Transp. Lett. **2**(3), 157–173 (2010). https://doi.org/10.3328/TL.2010.02.03.157-173
45. M. Treiber, A. Hennecke, D. Helbing, Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E **62**(2), 1805–1824 (2000). https://doi.org/10.1103/PhysRevE.62.1805
46. A. Kesting, M. Treiber, D. Helbing, General lane-changing model MOBIL for car-following models. Transp. Res. Rec. **1999**(1), 86–94 (2007). https://doi.org/10.3141/1999-10
47. J. Erdmann, Lane-changing model in sumo. Proc. SUMO2014 Model. Mobil. Open Data **24**, 77–88 (2014)
48. J. Wang, T. Shi, Y. Wu, L. Miranda-Moreno, L. Sun, Multi-agent graph reinforcement learning for connected automated driving, in *Conference: ICML Workshop on AI for Autonomous Driving* (2020), p. 7
49. S. Han, H. Wang, Stable and efficient Shapley value-based reward reallocation for multi-agent reinforcement learning of autonomous vehicles, in *2022 IEEE International Conference on Robotics and Automation* (2022)
50. B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, Y.P. Fallah, Social Coordination and Altruism in Autonomous Driving. IEEE Trans. Intell. Veh. (2022). https://doi.org/10.1109/TITS.2022.3207872
51. B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, Y.P. Fallah, Cooperative autonomous vehicles that sympathize with human drivers, in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2021), pp. 4517–4524. https://doi.org/10.1109/IROS51168.2021.9636151
52. B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, Y.P. Fallah, Altruistic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)* (2021)
53. D. Chen, Z. Li, M. Hajidavalloo, K. Chen, Y. Wang, L. Jiang, Y. Wang, Deep Multi-agent Reinforcement Learning for Highway On-Ramp Merging in Mixed Traffic (2022). arXiv:2105.05701 [cs, eess]
54. W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, D. Rus, Social behavior for autonomous vehicles. Proc. Natl. Acad. Sci. **116**(50), 24972–24978 (2019)
55. R. Valiente, B. Toghi, R. Pedarsani, Y.P. Fallah, Robustness and adaptability of reinforcement learning-based cooperative autonomous driving in mixed-autonomy traffic. IEEE Open J. Intell. Transp. Syst. **3**, 397–410 (2022)
56. W. Zhou, D. Chen, J. Yan, Z. Li, H. Yin, W. Ge, Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. Auton. Intell. Syst. **2**(1), 5 (2022). https://doi.org/10.1007/s43684-022-00023-5
57. Y. Hu, A. Nakhaei, M. Tomizuka, K. Fujimura, Interaction-aware decision making with adaptive strategies under merging scenarios, in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE Press, New York, 2019), pp. 151–158
58. C. Yu, X. Wang, X. Xu, M. Zhang, H. Ge, J. Ren, L. Sun, B. Chen, G. Tan, Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs. IEEE Trans. Intell. Transp. Syst. **21**(2), 735–748 (2020). https://doi.org/10.1109/TITS.2019.2893683
59. S. Bhalla, S. Ganapathi Subramanian, M. Crowley, Deep multi agent reinforcement learning for autonomous driving, in *Canadian Conference on Artificial Intelligence* (Springer, Berlin, 2020), pp. 67–78. https://doi.org/10.1007/978-3-030-47358-7_7
60. J. Foerster, I.A. Assael, N. De Freitas, S. Whiteson, Learning to communicate with deep multi-agent reinforcement learning. Adv. Neural Inf. Process. Syst. 29, (2016). https://doi.org/10.5555/3157096.3157336
61. S.K.S. Nakka, B. Chalaki, A.A. Malikopoulos, A multi-agent deep reinforcement learning coordination framework for connected and

automated vehicles at merging roadways, in *2022 American Control Conference (ACC)* (IEEE, New York, 2022), pp. 3297–3302
62. L. Wang, Z. Yang, Z. Wang, Breaking the curse of many agents: provable mean embedding q-iteration for mean-field reinforcement learning, in *International Conference on Machine Learning* (PMLR, 2020), pp. 10092–10103
63. L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, K. Kavukcuoglu, Impala: scalable distributed deep-rl with importance weighted actor-learner architectures, in *International Conference on Machine Learning*, vol. 80 (PMLR, 2018), pp. 1407–1416
64. J. Garcıa, F. Fernández, A comprehensive survey on safe reinforcement learning. J. Mach. Learn. Res. **16**(1), 1437–1480 (2015)
65. T. Özkan, T. Lajunen, J.E. Chliaoutakis, D. Parker, H. Summala, Cross-cultural differences in driving behaviours: a comparison of six countries. Transp. Res., Part F Traffic Psychol. Behav. **9**(3), 227–242 (2006)
66. E. Vinitsky, R. Köster, J.P. Agapiou, E. Duéñez-Guzmán, A.S. Vezhnevets, J.Z. Leibo, A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings (2021). arXiv preprint. arXiv:2106.09012
67. S.J. Grimbly, J. Shock, A. Pretorius, Causal Multi-Agent Reinforcement Learning: Review and Open Problems (2021). arXiv:2111.06721 [cs, stat]