# Safe Lane-Changing in CAVs Using External Safety Supervisors: A Review

Lalu Prasad Lenka(✉) and Mélanie Bouroche

Trinity College Dublin, Dublin, Ireland
{lenkal,melanie.bouroche}@tcd.ie

**Abstract.** Connected autonomous vehicles (CAVs) can exploit information received from other vehicles in addition to their sensor information to make decisions. For this reason, their deployment is expected to improve traffic safety and efficiency. Safe lane-changing is a significant challenge for CAVs, particularly in mixed traffic, i.e. with human-driven vehicles (HDVs) on the road, as the set of vehicles around them varies very quickly, and they can only communicate with a fraction of them. Many approaches have been proposed, with most recent work adopting a multi-agent reinforcement learning (MARL) approach, but those do not provide safety guarantees making them unsuitable for such a safety-critical application. A number of external safety techniques for reinforcement learning have been proposed, such as shielding, control barrier functions, model predictive control and recovery RL, but those have not been applied to CAV lane changing.

This paper investigates whether external safety supervisors could be used to provide safety guarantees for MARL-based CAV lane changing (LC-CAV). For this purpose, a MARL approach to CAV lane changing (MARL-CAV) is designed, using parameter sharing and a replay buffer to motivate cooperative behaviour and collaboration among CAVs. This is then used as a baseline to discuss the applicability of the state-of-the-art external safety techniques for reinforcement learning to MARL-CAV. Comprehensive analysis shows that integrating an external safety technique to MARL for lane changing in CAVs is challenging, and none of the existing external safety techniques can be directly applied to MARL-CAV as these safety techniques require prior knowledge of unsafe states and recovery policies.

**Keywords:** Connected autonomous vehicles · Lane changing · Multi-agent reinforcement learning · Safe reinforcement learning

## 1 Introduction

Improvements in traffic mobility, eco-friendly vehicles, fewer fossil fuels and safer roads are among the benefits promised by the autonomous vehicle industry. An autonomous vehicle can be defined as a vehicle capable of travelling unassisted anywhere and at any time, with no restrictions, and without the help or even the


L. Longo and R. O'Reilly (Eds.): AICS 2022, CCIS 1662, pp. 527–538, 2023.
https://doi.org/10.1007/978-3-031-26438-2_41

presence of a driver [1]. A connected autonomous vehicle (CAV) is an autonomous vehicle that can communicate with other vehicles or roadside units [1].

Though there has been great advancement in AV technology over the past decade, the number of traffic accidents involving autonomous vehicles has increased in recent years [2]. A major problem with automated driving at its current stage of development is that it is not yet reliable and safe [3].

As it is such a complex problem, a lot of the recent approaches to (connected) autonomous driving rely on artificial intelligence, and reinforcement learning (RL) in particular. RL agents, however, might visit unsafe states during the exploration phase, and there is no guarantee that they will not explore them during the exploitation phase.

This paper first reviews existing work in RL-based lane changing, safety approaches for lane changing and safety approaches for MARL more generally (Sect. 2). Recent papers addressing safe lane-changing of connected autonomous vehicles and safe multi-agent reinforcement learning were selected for review. It then presents a MARL design of lane changing for CAVs, building on the state of the art (Sect. 3). This design is used as a basis to discuss open challenges in safe MARL CAV lane changing (Sect. 4). Finally, the Sect. 5 concludes the paper.

## 2 Background

This section discusses the state-of-the-art lane changing approaches for CAVs (Sect. 2.1), before presenting the existing approaches to introduce safety for CAV lane changes (Sect. 2.2). Finally, it analyses the safety approaches for MARL (Sect. 2.3) applied in recent research.

### 2.1 RL-Based Lane Changing Approaches for CAV

Lane changing remains a challenging task where an autonomous vehicle has to predict the actions of neighbouring vehicles while changing lanes to avoid a collision. Furthermore, an autonomous vehicle is supposed to undertake the lane change manoeuvring without aggressive acceleration or deceleration to ensure the comfort of the passengers. The complexity of this problem has led many recent approaches to adopt a RL approach to lane changing. These RL-based autonomous vehicles controllers are dominated by the actor-critic methods, and the Deep Q Networks (DQN) and Deep Deterministic Policy Gradient (DDPG) algorithms [4]. Deep RL approaches are favoured over normal RL methods for high-dimensional state spaces. Policy-based methods are preferred in multi-agent setting. Ye et al. [7] proposed an automated mandatory lane change strategy using proximal policy optimization-based deep reinforcement learning, showing great advantages in learning efficiency and performance compared to Q-learning-based RL approaches.

One way of enhancing safety in multi-agent RL is through introducing cooperative behaviour and collaboration among agents. Fu et al. [5] modelled the lane-changing problem as a deep reinforcement learning process to learn the

optimal lane-changing strategy through a deep deterministic policy gradient (DDPG) algorithm. They also proposed a collective learning framework to use the collective intelligence of CAVs to improve the performance of autonomous lane-changing strategies. Zhou et al. [6] proposed a decentralized cooperative multi-agent reinforcement learning algorithm with an actor-critic policy for lane changing among CAVs, where they used parameter-sharing scheme to foster inter-agent collaborations.

Learning an optimal RL policy for lane changing can be challenging in the presence of mixed traffic where human driven vehicles (HDVs) are present along with controlled agents. Chen et al. used parameter sharing and local rewards to foster inter-agent cooperation and achieved high scalability in mixed traffic scenarios [8].

### 2.2 Safety Approaches in RL-Based Lane Changing for CAV

In the recent literature, most RL-based lane changing approaches introduced safety into the algorithm by modifying reward function, to include some parameters related to headway distance, collision avoidance, and passenger comfort.

Fu et al. discussed a blockchain-based collective learning framework for lane-changing in CAVs and introduced the headway distance (i.e. the distance between ego-vehicle and leading vehicle) into the reward function to make the lane changes safer [5]. Ye et al. introduced collision reward to penalize unsafe actions, and used a safety intervention module to label the output action from the algorithm as "catastrophic" or "safe" [7]. However, the details of the safety intervention module were not discussed.

Zhou et al. proposed the use of multi-agent actor-critic RL for cooperative lane changing among autonomous vehicles using the headway distance and a collision penalty in the reward function to add safety into their design [6]. Chen et al. [8] also used the headway distance and a collision penalty in the reward function to add safety, but also proposed a novel priority-based safety supervisor, which predicts the action of neighbouring vehicles to enable safer decisions. This is the first use of an external safety technique for CAV lane changing. However, though their approach is quite intuitive they do not discuss non-cooperative human driven vehicles.

Most papers discussed in this section focused on optimizing policies based on rewards and adding safety constraints to the reward function. But none truly guarantees safety, i.e., that no unsafe state is ever visited during the training and execution process.

### 2.3 Safety Approaches in MARL

Deep reinforcement learning techniques are able to maximize the intended reward, but they may not always ensure safety throughout the learning or execution stages. Safe Reinforcement Learning can be defined as the process of learning policies that maximize the expectation of the return, in problems in which it is

important to ensure reasonable system performance and/or respect safety constraints during the learning and/or deployment processes [10]. In the exploration phase the agent might encounter some unsafe states which makes Reinforcement Learning approach unsuitable for safety-critical systems as in this case failure can be costly [11].

According to García & Fernández [10] generally safety can be introduced into RL algorithm through two ways. The first way is to modify the reward function to include parameters for safety, so the safety is improved while the algorithm learns to optimize the reward function. The second approach is where the RL algorithm is modified to account for an external safety supervisor (such as teacher advice or demonstrations).

External safety approaches can be classified into two groups; the first type of technique uses some prior information about unsafe states to develop a safety critic that can provide the probability of agents being in unsafe situations in future states, and uses a separate recovery policy to bring the agents back to safe states. Model predictive control predicts the next states of each agent when the learnt policy is followed. If the predicted next states are recoverable for all agents, it uses the learnt policy. Otherwise it uses a recovery policy for the agents who will move into irrecoverable states after the current action [12,13]. Thananjeyan et al. proposed using a composite policy $\pi$ which selects between a task-driven policy and a recovery policy at each time step, based on whether the agent will violate safety constraints in the near future [11].

The Second type of techniques creates a set of safe states by estimating the dynamics of the system. They learn about unsafe states, project the actions of the RL agents into the safe set, and constrain unsafe actions. Control Barrier Functions (CBF) is a model-based safety framework that learns about safe states and prevents the exploration of dangerous states by projecting the RL agents' actions onto a safe set of actions. It has been used by a number of recent works [14–17] to introduce external safety in reinforcement learning tasks. ElSayed-Aly et al. specified safe states using finite state machines and used Linear Temporal Logic (LTL) as modal temporal logic to formally verify whether the visited state is a safe state [19].

Overall, according to our research there are four key external safety supervisor techniques used in the RL literature, i.e., shielding, control barrier functions, model predictive control, and recovery RL. This paper aims to study recent papers using these techniques and evaluate whether these approaches can be applied to improve the safety of lane changing in CAVs.

## 3 Modeling MARL-CAV

In this section, a decentralized MARL-based approach [20] for highway lane-changing of multiple CAVs is discussed. This approach is a **customized and adapted** version of the MARL ramp merging algorithm in Chen et al. [8]. It was modified for the lane changing scenario and a custom reward function was added (discussed below).

The mixed-traffic lane-changing environment is modelled as a multi-agent network: $\mathcal{G} = \{v, \varepsilon\}$, where each agent $i \in v$ communicates with neighbours $\mathcal{N}_i$ using the communication link $\epsilon_{ij} \in \varepsilon$. Here $\mathcal{N}_i$ represents the set of agents in close proximity to agent $i$. The overall dynamic system can be considered a partially-observable Markov decision process (POMDP) which can be represented by the tuple $(\{\mathcal{A}_i, \mathcal{S}_i, \mathcal{R}_i\}_{i \subseteq v}, \mathcal{T})$, where $\mathcal{A}_i$ is the local action space, $\mathcal{R}_i$ is the reward space, $\mathcal{O}_i \in \mathcal{S}_i$ is the partial observation of the environment state [6].

In partially observable Markov games (multi-agent POMDP), every agent follows a decentralized policy $\pi_i : \mathcal{O}_i X \mathcal{A}_i \rightarrow [0, 1]$ to chose its own action $a_{i,t} \sim \pi_i(\cdot|s_{i,t})$ at time step $t$. The action space $\mathcal{A}_i$ of agent i is defined as a set of high-level **discrete control decisions (actions)**, including: cruising, turn left, turn right, speed up and slow down. The state space $\mathcal{O}_i$ of contains the longitudinal and latitudinal features (speed and position). The reward function is designed to achieve the goal of safe lane changing. Our reward function is composedly designed using multiple metrics like safety, headway distance, driving speed, right lane driving, lane changing:

$$r_{i,t} = w_s r_s + w_h r_h + w_d r_d + w_{rl} r_{rl} + w_{lc} r_{lc} \tag{1}$$

where $w's$ are weighing coefficients and $r's$ are cost evaluation.

In the proposed approach a deep neural network is used to approximate the stochastic decentralized policy $\pi$ of the RL agents. This network is shared between all agents. In addition, a shared replay buffer is also maintained to store the experiences from all agents. A copy of the state information i.e. observations, actions and rewards, is held by the individual agents. No agent has access to the state information of any other agent. However, as the data in the Replay Buffer is shared and identical, each agent benefits from the collective experiences of all agents. Finally, each agent updates the policy network asynchronously at each step [9]. In this design, the MARL algorithm does not have access to information about unsafe states and recovery policies.

## 4 Using an External Safety Technique for MARL-CAV

The design of MARL-CAV discussed in Sect. 3 only uses the reward function to introduce some level of safety. This is achieved by penalising actions which lead to a collision and rewarding the agents when they maintain a suitable headway distance from the front vehicle. However, this is not sufficient to prevent the agents from visiting unsafe states especially during the exploration phase and subsequently in the exploitation phase. Hence, an external safety supervisor might be useful to stop the agent from visiting unsafe states.

An external safety supervisor can be applied to enhance the agents' safety. The safety approaches discussed in Sect. 2.3 namely shielding, control barrier functions, model predictive control, and recovery RL, should integrate with MARL-CAV and then we can examine whether these techniques can improve safety.

This section first discusses the safety requirements for lane changing in CAV (Sect. 4.1), and then uses them to analyse the suitability of the external safety techniques for MARL CAV. Finally, it discusses open challenges.

### 4.1 Safety Requirements for Lane Changing in CAV

AI approaches are often validated on simple scenarios. In contrast, lane changing is a challenging scenario. Not only it contains an unspecified and unbounded number of controlled agents, but also (an unspecified an unbounded) number of uncontrolled agents (e.g., human-driven vehicles). Also, to fit with our approach (Sect. 3), the safety approach should be compatible with discrete action space as this library only supports discrete actions: move to the left lane, move to the right lane, forward, idle.

The Table 1 shows the requirements specific to the MARL-CAV that the safety approach should satisfy. The next section analyses the applicability of the safety techniques to MARL-CAVs and examines whether they satisfy the requirements mentioned in the Table 1.

**Table 1.** Characteristics of Lane Changing in CAVs

| Characteristic | Safety approach requirement |
|---|---|
| Number of agents | Should support a **multi-agent** scenario. It should be scalable to a large number of agents |
| Mixed traffic scenario | Should support the presence of **both** controlled and uncontrolled agents in the environment |
| Action Space | Should support **discrete actions** as the current implementation of MARL-CAV has discrete actions |
| Unsafe states information | Should **not require prior information** about unsafe states as the simulation is very dynamic |
| Recovery/Backup Policies | Should **not require a recovery or backup policy** developed to get the agent from unsafe to safe states |

### 4.2 Analysis of External Safety Techniques for MARL

Section 2.3 showed that, generally, control-theoretic or formal methods like shielding, control barrier functions, model predictive control and recovery policies are used to introduce external safety in MARL implementation for different domains. This section presents a more detailed analysis of these external safety approaches.

**Shielding.** In the shielding method, finite state machines are used to model all possible states an agent can visit, and linear temporal logic (LTL) safety specifications are used to design shields that restrict the agent from visiting unsafe states. ElSayed-Aly et al. [19] claim that the shielding approach not only guarantees safety but also learns more optimal policies with better returns than non-shield MARL, as unsafe actions which may destabilize learning are removed. However, shielding requires prior knowledge of safe states in the environment so that they can be used to design shields. Safety is specified using Linear Temporal Logic (LTL). This might be possible for a simpler environment like navigation tasks in ElSayed-Aly et al. [19] but could be difficult for complex environments like lane changing in connected autonomous vehicles. As discussed in the Table 1, there is no prior information about unsafe states in the MARL-CAV design.

**Control Barrier Functions.** Control Barrier Functions (CBF) is a model-based safety framework that prevents the exploration of dangerous states by projecting the RL agent's actions onto a safe set of actions. With CBFs safety is specified by defining a forward invariant set in the state space in which the system is **required to stay**. Given a time-varying set $\mathscr{C} \subset \mathbb{R}^n$ defined as zero superlevel set of a continuously differentiable function $h : \mathbb{R}^n x \mathbb{R}_+ \rightarrow \mathbb{R}$

$$\mathscr{C} \triangleq \{x \in \mathscr{R} : h(x,t) \geq 0\} \tag{2}$$

where $\mathscr{C}$ is the safe set, $h$ is the CBF. The idea of forward invariance is that if we have a state, we want to make sure that the state of the system would stay inside the set for long time. Control Barrier Functions can be used as a formulation tool to achieve forward invariance and, therefore the safety of a set. This is a very promising idea and hence control barrier function is the most widely used safety technique out of the four techniques discussed.

As discussed in Sect. 2.3, several approaches to MARL have introduced external safety through CBFs. These papers have simulated different tasks such as navigation tasks, i.e. an agent aims to move from a source to destination like in [14,15], or any OpenAI tasks like car following environment [17], unicycle environment [17].

These approaches [15,18] used continuous action spaces as CBF would not work for a discrete action space. Control barrier functions do not work for discrete action space because with discrete action space the RL model cannot be represented using differential dynamic programming and the lie derivatives would not work. This is a major drawback as per the safety requirements discussion in Sect. 4.1.

A key problem for this approach is figuring out how to combine knowledge of the model (environment) dynamics with model-based safety, as control barrier functions are model-based i.e. they require information about model dynamics. Hence most of these papers either use model-based RL like in [17] or use a statistical model to learn model dynamics like in [15,18]. However, in POMDP discussed in Sect. 3 the system dynamics can be very uncertain. There is no guarantee that a complex setting of multi-agent RL for lane changing with mixed

traffic scenario can be described with some statistical assumptions of dynamics. Complex environment dynamics of MARL-CAV makes use of CBF challenging for lane changing in CAVs.

**Model Predictive Control.** Model predictive control can be defined as single or multi step ahead estimation of states that can be reached by an agent over multiple next time steps under a sequence of control inputs. Generally, statistical models are used to approximate the system dynamics, which helps to predict the successive states of the environment when an action is taken by the agent. The predicted states are then verified if they violate safety constraints and if an agent has a high probability of reaching unsafe states, it uses a recovery (or backup) policy that brings it back to safe states.

MPC can also be applied in a multi-agent setting, like in [12]. In their proposed method, they are incrementally checking whether each agent is in a set of stable states $\chi_{stable}$, and if any agent is predicted to be going into an irrecoverable or unsafe state, then recovery policy $\pi_{recovery}$ is used to bring the agent back to safe states $\chi_{stable}$. The agents predicted to be in safe set after current action use the respective learned policy $\pi^i$. Unlike our implementation of parameter sharing discussed in Sect. 3 they have assumed different learned policy for each agent which is not scalable to higher number of agents.

This approach looks promising and able to act as a safety supervisor for the RL agent. However, these approaches generally have two underlying assumptions. First, availability of a recovery policy, also called safe policy $\pi_{safe}$ is used to return the agent to safe state when it is about to visit the unsafe states. Second, some prior information of either the irrecoverable (unsafe) states or safe states.

The challenge is that the recovery policy and data of unsafe states for an environment might not be available for all use cases. For example, in a multi-agent lane changing the setting, it is difficult to know which all states are unsafe and how one trains a recovery policy. These challenges are limitations to using this safety technique for multi-agent lane-changing scenarios.

**Recovery RL.** The Recovery RL approach is relatively very recent compared to the previous three approaches. Thananjeyan et al. [11] proposed a unique approach to create an external supervisor for RL algorithms. However, it has some similarities to model predictive control methods, especially the one in [12].

According to Thananjeyan et al. [11], an offline dataset $D_{offline}$ was generated which contains set of trajectories where the agent violated safety constraints. This was either generated manually using human knowledge or through a naive RL policy. This $D_{offline}$ data was used to train a reinforcement learning-based safety-critic policy. The safety critic was later used to estimate the agent's probability of future constraint violations. If the safety-critic predicts the agent's action to be unsafe, then recovery policy $\pi_{recovery}$ is used to bring it back to safe states or else a learn policy $\pi_{task}$ is used.

This approach requires us to design the $D_{offline}$ either manually using human knowledge or through a separate reinforcement learning policy. For complex

systems like the multi-agent RL model for lane changing in CAVs, it would require the first training an RL policy and extracting the $D_{offline}$. Though the $D_{offline}$ would be updated during the training of the recovery RL it does require to run the RL model without the safety critic to collect $D_{offline}$. This in-turn defeats the purpose of designing an external safety supervisor, as the main aim of external safety is to stop the agent from exploring unsafe states.

The Table 2 provides the analysis of safety techniques concerning the safety requirements. The green color cell means the safety method satisfies safety requirements. From the table, it can be inferred that none of the safety techniques satisfies all the safety requirements; hence, there are still some open challenges in using them for MARL-CAV. These challenges are discussed in more detail in the next section.

**Table 2.** Analysis of the External Safety Supervisors for MARL in terms of the scenario requirements. Green color cell means scenario requirement is satisfied

| **Characteristic** | **Shielding** | **Control Barrier Functions** | **Model Predictive Control** | **Recovery RL** |
|---|---|---|---|---|
| **Number of agents** | Supports multi-agent | Supports multi-agent | Supports multi-agent | Supports multi-agent |
| **Mixed agents (controlled & uncontrolled)** | No support yet | No support yet | No support yet | No support yet |
| **Action Space** | Continuous | Continuous | Continuous | Continuous and Discrete |
| **Model dynamics information** | Required apriori | Modelled using Gaussian Processes or Assumed | Modelled using Gaussian Processes, or Assumed | Learned using RL |
| **Unsafe states information** | Required | Not required | Required | Required |
| **Recovery/Backup Policies** | Not required | Not required | Required | Required |

### 4.3 Discussion

As discussed in Sect. 2.2, most MARL methods for CAV modelling mostly focus on achieving optimal policies based on their reward function but this does not guarantee safety, i.e., that no unsafe state is ever visited during the learning process. This is because penalizing agents through negative rewards does not stop them from going to unsafe states as they have to explore different (both safe and unsafe) actions to learn about the states and rewards from the environment.

The approaches discussed in Sect. 4.2, namely shielding, model predictive control, control barrier functions and recovery RL, are ways to guarantee safety during the exploration process as they act like an external supervisor, thereby monitoring the agents' actions and preventing the exploration of unsafe states.

These approaches have been used to provide external safety for MARL for other scenarios and might be applicable to the multi-agent reinforcement learning algorithms for lane changing in CAVs. The multi-agent proximal policy optimization algorithm has to be modified to include one of these safety approaches.

However, this implementation will be a significant challenge. Based on the data in Table 2 it can be inferred that most of these approaches are applied to simpler tasks, such as one or more agents trying to reach a goal [14,15], navigation tasks [11] or simple OpenAI gym environments [17]. These approaches use statistical models to estimate the dynamics of the system, but this might not work for a complex scenario like lane-changing in CAVs, where other human-driven vehicles, i.e., uncontrolled agents, are present. Compared to the static obstacles present in simpler environments, these represent dynamic obstacles. Furthermore, as discussed in Sect. 4.2, these approaches come with several assumptions such as:

- Availability of partial or complete prior knowledge of unsafe or irrecoverable states in shielding, model predictive control and recovery RL.
- Availability of recovery policies in the case of model predictive control and recovery RL.
- Bounded system dynamics for control-theoretic methods like control barrier functions.
- Having only a single or more controlled agents in the system. Absence of uncontrolled agents (like human-driven vehicles in the lane-changing scenarios).

As per the safety requirements in Sect. 4.1, these assumptions will be violated in the MARL-CAV setting discussed in Sect. 3 and this posses a significant challenge in adapting these approaches to MARL-CAV. Based on the assumptions and limitations discussed in this chapter, the potential introduction of these safety approaches to MARL-CAV comes with theoretical and practical challenges.

## 5 Conclusion

The investigation of recent literature showed that most of the state-of-the-art implementations of multi-agent reinforcement learning (MARL) for lane changing in CAVs use a custom reward function to introduce safety, but this is not completely safe as the agents still visit unsafe states during the exploration phase of the RL algorithm. The characteristics of safe MARL for lane changing scenarios were discussed (summarised in Table 1) and state-of-the-art external safety techniques for MARL were researched and analysed. It was inferred that an ideal safety supervisor approach should support both continuous and discrete actions, should operate in a mixed traffic scenario, should not require a recovery policy or any prior information about unsafe states. Based on these, the functioning and limitations of different external safety supervisors for multi-agent RL were analysed and summarised in Table 2. It was observed that these safety techniques

come with many assumptions like availability of recovery policies, prior information about unsafe states, the requirement of continuous action space etc. and are often not designed for mixed agents (controlled and uncontrolled) scenarios.

These assumptions mean these techniques cannot be directly applied to MARL for CAVs. Integration of these safety approaches to MARL-CAV is not straightforward and requires significant work, which involves modifying either the MARL-CAV implementation or the safety techniques. In future work we aim to modify the design of MARL-CAV and make tweaks in safety supervisor techniques to enable successful integration of the safety techniques with MARL-CAV.

## References


1. Paret, D., Rebaine, H., Engel, B.A.: The buzz about autonomous and connected vehicles, pp. 3–22. Wiley (2022)
2. Dixit, V.V., Chand, S., Nair, D.J.: Autonomous vehicles: disengagements, accidents and reaction times. PLOS One **11**(12), e0168054 (2016)
3. Martens, M., van den Beukel, A.: The road to automated driving: dual mode and human factors considerations. In: ITSC 2013, pp. 2262–2267 (2013)
4. Haydari, A., Yılmaz, Y.: Deep reinforcement learning for intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. **23**(1), 11–32 (2022)
5. Fu, Y., Li, C., Yu, F.R., Luan, T.H., Zhang, Y.: An autonomous lane-changing system with knowledge accumulation and transfer assisted by vehicular blockchain. IEEE Internet Things J. **7**(11), 11123–11136 (2020)
6. Zhou, W., Chen, D., Yan, J., Li, Z., Yin, H., Ge, W.: Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. Auton. Intell. Syst. **2**(1), 5 (2022)
7. Ye, F., Cheng, X., Wang, P., Chan, C.-Y., Zhang, J.: Automated lane change strategy using proximal policy optimization-based deep reinforcement learning. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1746–1752 (2020)
8. Chen, D., et al.: Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic (2022). https://doi.org/10.48550/arXiv.2105.05701
9. Kaushik, M., Singhania, N., Krishna, K.M.: Parameter sharing reinforcement learning architecture for multi agent driving. In: AIR 2019, Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3352593.3352625
10. García, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. J. Mach. Learn. Res. **16**(1), 1437–1480 (2015)
11. Thananjeyan, B., et al.: Recovery RL: safe reinforcement learning with learned recovery zones (2020). https://arxiv.org/abs/2010.15920
12. Zhang, W., Bastani, O., Kumar, V.: MAMPS: safe multi-agent reinforcement learning via model predictive shielding (2019). https://arxiv.org/abs/1910.12639
13. Zanon, M., Gros, S.: Safe reinforcement learning using robust MPC. IEEE Trans. Autom. Control **66**(8), 3638–3652 (2021)
14. Cai, Z., Cao, H., Lu, W., Zhang, L., Xiong, H.: Safe multi-agent reinforcement learning through decentralized multiple control barrier functions (2021)
15. Qin, Z., Zhang, K., Chen, Y., Chen, J., Fan, C.: Learning safe multi-agent control with decentralized neural barrier certificates (2021)

16. Ames, A.D., Xu, X., Grizzle, J.W., Tabuada, P.: Control barrier function based quadratic programs for safety critical systems. IEEE Trans. Autom. Control **62**(8), 3861–3876 (2017)
17. Emam, Y., Glotfelter, P., Kira, Z., Egerstedt, M.: Safe model-based reinforcement learning using robust control barrier functions (2021). https://arxiv.org/abs/2110.05415
18. Zhao, H., Zeng, X., Chen, T., Liu, Z., Woodcock, J.: Learning safe neural network controllers with barrier certificates. Formal Aspects Comput. **33**(3), 437–455 (2021). https://doi.org/10.1007/s00165-021-00544-5
19. ElSayed-Aly, I., Bharadwaj, S., Amato, C., Ehlers, R., Topcu, U., Feng, L.: Safe multi-agent reinforcement learning via shielding. In: AAMAS 2021, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, pp. 483–491 (2021)
20. Leurent, E.: An environment for autonomous driving decision-making (2018). https://github.com/eleurent/highway-env