Learning-Based Social Coordination to Improve Safety and Robustness of Cooperative Autonomous Vehicles in Mixed Traffic

Rodolfo Valiente, Behrad Toghi, Mahdi Razzaghpour, Ramtin Pedarsani, and Yaser P. Fallah

1 Introduction

Autonomous vehicles (AVs) have been an attractive research area for decades, as it offers the potential to generate more efficient and safer road networks [1]. The adoption of AVs will not become a reality until they can co-exist with humans, as part of a complex social system. In order to maximize the potential of AVs and optimize for safety and traffic efficiency of all the vehicles on the road, AVs have to coordinate and influence the other agents [1–3].

We recognize the importance of social interaction and behavior in safety and reliability and identify two important research directions. First, AVs must be social actors and behave predictably and safely. Driver behavior is shaped by habits and expectations in the traffic environment. The vehicle's interaction will be influenced by the way AV decisions are perceived. Therefore, the ability of AVs to drive in a socially obedient manner is critical for the safety of passengers and other vehicles because predictable behavior allows humans to comprehend and respond appropriately to the AV's actions. Second, AVs must be social-aware and learn to identify social cues of egoism or altruism, understand the behavior of human drivers and learn how to interact and coordinate with all agents in a mixed traffic

R. Valiente (✉) · Behrad Toghi · Mahdi Razzaghpour · Y. P. Fallah
Connected & Autonomous Vehicle Research Lab (CAVREL), University of Central Florida, Orlando, FL, USA
e-mail: rvalienter90@knights.ucf.edu; toghi@knights.ucf.edu; razzaghpour.mahdi@knights.ucf.edu; Yaser.Fallah@ucf.edu

R. Pedarsani
Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA
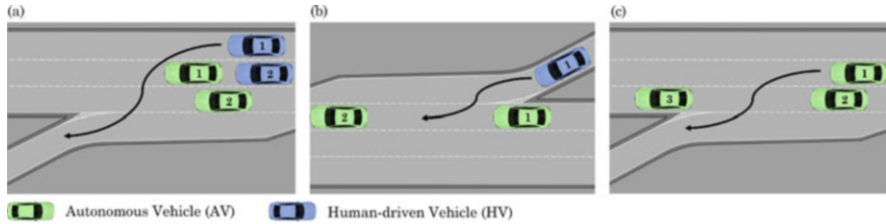e-mail: ramtin@ece.ucsb.edu

Fig. 1 **(a) Interaction of AV-HV to benefit a HV:** Altruistic agents create alliances and direct the behavior of HVs to improve traffic flow and prevent dangerous circumstances. AV1 and AV2 can create a formation to guide HV2 and provide a route for HV1, allowing the HV to change lanes and navigate to the exit ramp. **(b) Interaction of AV-AV to benefit a HV:** The goal of HV1 is to integrate onto the highway. Egoistic AVs disregard the merging vehicle and do not make room for it, possibly resulting in dangerous situations, however, if they exhibit sympathy for the merging HV, they can compromise on their own interest to create a safe path for HV1 to merge into the highway. **(c) Interaction of AV-AV to benefit another AV:** The goal of AV1 is to exit the highway. If AV2 acts selfishly, AV1 may miss the exit and be unable to complete its task. However, if AV2 and AV3 consider AV1's mission and act altruistically, they can free up space in the platoon by AV2 decelerating and AV3 accelerating, allowing AV1 to safely take the exit

environment, adapting and influencing the HVs behaviors to optimize for a social utility that improves traffic flow and safety.

In this chapter, we focus on social awareness challenges and seek a solution that can ensure the safety and robustness of AVs in the presence of human drivers with heterogeneous behavioral traits. Vehicle-to-vehicle (V2V) communication allows connected and autonomous vehicles (CAVs) to interact directly with their neighbors [4, 5]. By using V2V communication CAVs can create an extended perception that facilitates explicit cooperation among vehicles to overcome the limits of a non-cooperative agent [6–8]. While planning in a fully AV scenario is relatively easy to achieve, coordination in the presence of HVs is a significantly more challenging task, as the AVs not only need to react to road objects but also need to consider the behaviors of HVs [9–11].

In contrast to the individual non-cooperative approaches, we investigate the mixed-autonomy decision-making challenge from a multi-agent and social perspective. Rewarding AVs for adopting an altruistic behavior and taking into consideration the interests of other vehicles allows them to see the broad picture and find solutions that maximize the utility of the group. In addition to the potential benefits of altruistic decision-making in terms of safety and efficiency, altruism results in more societally advantageous outcomes [2]. Figure 1a demonstrates how a group of AVs can guide HV to increase safety and efficiency, while Fig. 1b and c show how AVs can collaborate to accomplish a social objective that benefits another HV or AV.

Currently, AVs lack an understanding of human behavior and frequently act extremely cautiously to avoid collisions. This conservative behavior not only leaves AVs unprotected from aggressive HVs, but also results in unexpected reactions that confuse HVs, creating bottlenecks in traffic flow and causing accidents. It's critical

to distinguish between a human driver's individual traits, such as aggressiveness, conservativeness, and risk tolerance, and their social preferences, such as egoism and altruism. Even though the two categories are related, they have distinctive natures and so behave differently in mixed traffic. An aggressive driver, for example, is not inherently egoistic or selfish, but their aggression may hinder their ability to collaborate with other drivers and participate in a socially desirable coexistence with AVs [3, 12, 13]. In the field of behavior planning for AVs in mixed-autonomy traffic, we identify two fundamental problems. First, human drivers differ considerably in their individual traits and social preferences, making AV behavior planning exceedingly difficult because it is difficult for the AV to foresee the type of behavior it would encounter when dealing with a human driver. Furthermore, relying on a real-time inference of HV behaviors is not always feasible because vehicle interactions can be brief, such as when two vehicles meet at an intersection. Second, driving requires complex interactions of agents in a partially observable and non-stationary environment, as HVs do not follow a fixed policy and modify their policies in real-time in response to the actions of other vehicles.

The integration of AVs into the real world requires them to address those challenges. Due to the differences in maneuverability and reaction time between AVs and HVs, a road shared by both becomes a competitive situation. In contrast with the full-autonomy case, here the coordination between HVs and AVs is not as straightforward since AVs do not have an explicit means of harmonizing with humans and are therefore required to locally account for the other HVs and AVs in their proximity. This dilemma intensifies if AVs act egoistically and optimize solely for their local utility. As an illustration, Figs. 2 and 3 demonstrate a highway exiting and merging scenario in mixed traffic. We consider a general setup where AVs and HVs with different behaviors coexist. Vehicles need to efficiently merge onto the lane or exit the highway without collisions. In an ideal cooperative environment, AVs should proactively decelerate or accelerate to provide sufficient room for vehicles to safely exit/merge and prevent hazardous situations, while also being resilient to various situations and behaviors and assuring safety in decision-making [14]. For instance, in Fig. 2 (merging scenario) if AVs act egoistically, the merging vehicle must rely on the HV to slow down to allow it to merge. However, due to the unpredictability of HVs, relying solely on HVs might result in suboptimal or even dangerous circumstances. Therefore, if all AVs act egoistically, the merging vehicle would either be unable to join the highway or it will wait for an HV and risk cutting into the highway without knowing whether the HV will slow down. Nevertheless, if AVs act altruistically, they can coordinate to guide the traffic on the highway to allow for a seamless and safe merging. In particularly challenging driving scenarios, such altruistic AVs can achieve societally beneficial results without relying on or making assumptions about HVs behaviors.

To address these challenges, existing literature either depend on models of human behavior generated from pre-recorded driving datasets [15, 16] or define social utilities that can impose cooperative behavior among AVs and HVs [17]. Other works focus on rule-based methods that use heuristics and hand-coded rules to guide the AVs [18] or probabilistic driver modeling [19] learned from human driving data.
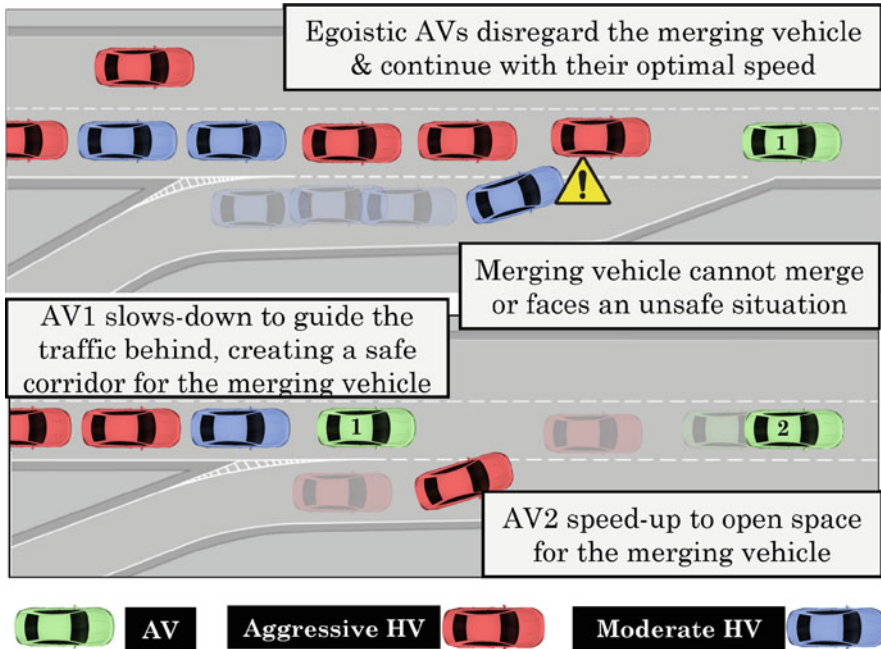
Fig. 2 For a seamless and safe highway merging, all AVs must coordinate and account for the utility of HVs. *(top)* Egoistic AVs optimize only for their own utility, *(bottom)* Altruistic AVs consider also the HV's utility

While this is feasible for simple situations, these methods become impractical in complex scenarios. Additionally, the human driver models learned in the absence of AVs, are not necessarily valid when humans confront AVs. This limits the application of the generated solutions, as they are frequently limited to the human behaviors with which AVs interacted during training. To account for this, several works in the literature adopt an excessively cautious approach when interacting with humans [20]. This strategy not only leaves the AVs vulnerable to other aggressive drivers, particularly in competitive situations, but it also causes traffic congestion and significant safety risks [1, 2].

On the other hand, data-driven methods such as reinforcement learning (RL) have received increased attention [21] as RL-based methods can learn decision-making and driving behaviors that are hard for traditional rule-based designs. However, the majority of the RL approaches are designed for a single AV, or try to handle the interaction between AVs and HVs either by predicting human behavior or by relying on the fact that humans are willing to collaborate or can be influenced to do so [15, 22], which could compromise safety or lead to sub-optimal performance. Recent works consider social interactions of AVs and train altruistic AVs that learn from experience and influence HVs to optimize a social utility function that benefits all vehicles on the road [10, 23].
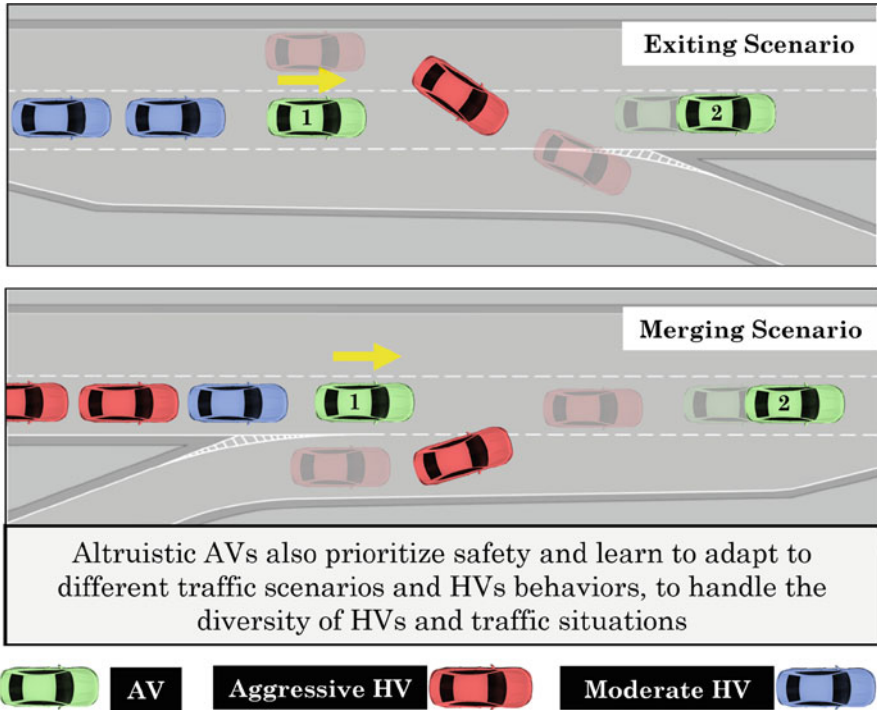
Fig. 3 Highway exiting and merging scenarios with AVs (green) and aggressive HVs (red) sharing the road. Altruistic AVs must learn to cooperate to exit/merge successfully and safely while being adaptable to a variety of scenarios and HV behaviors

In contrast, we consider a data-driven multi-agent reinforcement learning (MARL) approach and let the autonomous agents implicitly learn the decision-making process of human drivers only from experience, while optimizing for a social utility. By incorporating a cooperative reward structure into our MARL framework, we can train AVs that coordinate with each other, sympathize with HVs, and, as a result, demonstrate enhanced performance in competitive driving scenarios, such as highway exiting and merging. Despite not having access to an explicit model of the human drivers, the trained autonomous agents learn to implicitly model the environment dynamics, including the behavior of human drivers, which enables them to interact with HVs and guide their behavior.

This research aims to create a safe and robust training regimen that allows AVs to collaborate and influence the behavior of human drivers to achieve socially desirable outcomes, regardless of HV individual traits and social preferences. We based our work on the following insights. First, we rely on a decentralized reinforcement learning architecture that optimizes for a social utility that learns from experience and exposes the learning agents to a wide range of driving. As a result, the agents become more resistant to human driver behavior and can handle cooperative-

Fig. 4 An overview of our approach to leverage social awareness and coordination to improve the safety and reliability of CAVs. Our social-aware AVs learn from scratch not only to drive but also to understand the behavior of HVs and coordinate with them, they learn to adapt and influence HVs in a robust and safe manner

competitive behaviors regardless of HV's hostility or social preference. Second, a safety prioritizer is presented to minimize high-risk actions that could jeopardize driving safety. The safety prioritizer constrains the policy of cooperative AVs to ensure the safety of their behavior via masking the Q-states that lead to high-risk outcomes. Figure 4 shows an overview of our process.

Our main contributions are summarized as follows:

- We formulate the mixed-autonomy problem as a decentralized MARL problem and present an approach to training altruistic agents which utilizes a decentralized reward mechanism for achieving socially advantageous behaviors and takes advantage of a 3D convolutional deep reinforcement learning architecture to capture the temporal information in driving data.
- A training algorithm is proposed to make AVs robust to different drivers' behavior and situations while producing socially desirable outcomes. We investigate the effect of HVs behaviors on our altruistic AVs agents and especially conclude that the higher the traffic aggressiveness, the higher the importance of social coordination.
- We investigate the scenarios in which altruistic AVs can learn cooperative policies that are robust to diverse traffic scenarios and HV behaviors without compromising efficiency and safety, and present the results on transfer learning and domain adaptation in mixed-autonomy traffic.

The purpose of this chapter is to study the challenges of robust and safe AVs in mixed-autonomy traffic, especially in intrinsically competitive driving scenarios like those shown in Fig. 2, in which coordination is essential for safety and efficiency. The intention is to utilize the autonomous driving challenge as a case study to examine the use of social theories from psychology literature in the MARL domain. To apply these theories to real-world roads, more study is required. Nonetheless, the research on altruistic AVs that are robust, safe, and capable of learning to influence HVs in desirable ways, without the limitations of current solutions are promising.

2 Related Work

2.1 Multi-Agent Reinforcement Learning

The intrinsic non-stationarity of the environment is a key problem for MARL. To address those limitations a MARL derivation of importance sampling is proposed and used to remove the outdated samples from the replay buffer [24]. In [25] is presented another solution to address this issue by including latent representations of partner strategies allowing partner modeling and more scalable MARL.

To mitigate the problem of credit assignment in multi-agent systems, [26] proposed the *counterfactual multi-agent (COMA)* algorithm, which employs a centralized critic and decentralized actors. In [27] is proposed a deep RL algorithm with full environment observability and a centralized controller to govern the joint-actions of all the agents. Other current research on mixed-autonomy focuses on addressing cooperative and competitive challenges by assuming the nature of interactions between autonomous agents [28]. In [29] a variation of an actor-critic approach with a *centralized q-function* is proposed. The algorithm has access to local observations and the actions of all agents. In our work, in contrast, we consider a decentralized controller with partial observability, and train altruistic agents that optimize for a social utility.

2.2 Driver Behavior and Social Coordination

Existing works on driver behavior and social navigation approach agents coordination by either modeling driver behaviors [19, 30, 31] or simplifying and making assumptions about the nature of agent interactions [28, 32]. In [33] is presented a maneuver-based dataset and a model for classifying driving maneuvers is proposed. Other works on driver behavior modeling consider graph theory [34], data mining [35], driver attributes [36] or game theory [2]. In [31] is proposed a method for modeling and forecasting human behavior in circumstances that involve multi-human interactions in highly multi-modal situations.

Current research in social navigation has demonstrated the importance of AVs as social actors and the advantages of coordination between AVs and HVs [37]. Human driving patterns are learned from demonstration using inverse RL in [38] and [22]. Similarly, in [39] is presented a centralized game-theory model for cooperative inverse RL. The authors in [40] and [41] proposed a shared reward function to enable cooperative trajectory planning for robots and humans. Sadigh et al. presents a strategy based on imitation learning to learn a reward function for human drivers, demonstrating how AVs can influence human actors [15]. The importance of coordination and the advantage of using AVs to guide the traffic has been also investigated at the traffic level. Wu et al. [42] analyzes the capability of AVs to stabilize a system of HVs and presents the conditions in which when concurrently

enforcing safety constraints on the AVs while stabilizing traffic improves traffic performance. Similar works have highlighted the potential of influencing HVs and how AVs can be used to stabilize and guide the traffic flow [42, 43]. Recent works focus on optimizing traffic networks in mixed autonomy to reduce traffic congestion and improve safety. In [44] is presented a model of vehicle flow and a model of how AV makes decisions among routes with various prices and latencies. The planner optimizes for a social objective and shows improvement in traffic efficiency. The vehicle routing problem is studied in [45] that proposed an innovative learning-augmented local search system to mitigate the problem by using a Transformer architecture. Cameron et al. explores how humans can supervise agents in order to attain an acceptable degree of safety [46]. In contrast to previous works, we do not rely on human cooperation and our AVs learn cooperative behaviors directly from experience, our focus is on the emerging altruistic behavior that allows agents to coordinate and optimize for a social utility.

2.3   Safe and Robust Driving

Safety is critical for AVs [47], and it is especially important for AVs that have been trained via RL. We must prioritize safety; because coordination is frequently associated with risk. In cooperative driving, there are often safe actions that have low rewards and riskier actions with higher rewards [48]; however, the risky action increases the likelihood of crashes when cooperation fails. Especially, AVs utilizing trained RL algorithms, may not always operate safely since the trained models may pick dangerous actions [20]. Several attempts in this direction use pure reward shaping to avoid collisions. While this is a frequent technique in RL, safety is not implicitly emphasized, and AVs implementing such RL methods may not behave properly in some cases due to function approximation.

To overcome this problem, the concept of safe RL is proposed in [20], which aims to increase safety in unobserved driving conditions when the RL algorithm performs dangerously. Wang et al. [49] proposes a rule-based decision-making system that evaluates the controller's decisions and substitutes collision-causing actions. A short-horizon safety supervisor is included in Nageshrao et al. [50] to replace unsafe actions with safer ones. A Q-masking strategy is presented in [51] to prevent collisions by deleting actions that might lead to a crash. Chen et al. proposes a novel priority-based safety supervisor that reduces collisions considerably [52].

We leverage these approaches in this work using a decentralized reward function, local actions, and assuming partial observability, to increase the altruistic agents' safety while also being adaptable to varied driver behaviors and circumstances. As shown in Fig. 2, we analyze a particular situation in which AVs and HVs with various characteristics coexist. The picture depicts two frequent traffic situations in which vehicles must either merge into a lane effectively or depart the highway without colliding with other vehicles. In an ideal cooperative context, vehicles should proactively decelerate or accelerate to provide enough room for vehicles

to safely exit/merge and prevent stalemate situations, while also being resilient to various conditions and behaviors and assuring safety in decision-making.

3   Preliminaries and Formalism

We study safety and robustness in the maneuver-level decision-making problem for AVs to see what kinds of behaviors might lead to socially desirable results. We're interested in the question of how AVs can be trained from scratch to drive safely and reliably, while also taking into account the social aspects of their mission, i.e., optimizing for a social utility that takes into account the interests of other vehicles in the vicinity. Social awareness and coordination are essential to improve safety and reliability on the roads. In this work, we explore that insight. Thus, we continue this section by providing a quantitative description of an agent's level of altruism and formally defined our problem.

It is possible to define the MARL problem as a centralized or decentralized problem. It's simple to create a centralized controller that provides a central joint reward and joint action. However, in the real world, such assumptions are unfeasible. In this chapter, we focus on a decentralized controller with partial observability and formulate the problem as a partially observable stochastic game (POSG) defined by $\langle \mathcal{I}, \mathcal{S}, P, \gamma, \{\mathcal{A}_i\}_{i \in \{1,...,N\}}, \{O_i\}_{i \in \{1,...,N\}}, \{R_i\}_{i \in \{1,...,N\}} \rangle$ where

- $\mathcal{I}$: a finite set of agents $N \geq 2$.
- $\mathcal{S}$: a set of possible states that contains all configurations that N AVs can take (probably infinite).
- P: a state transition probability function from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$, $P(S = s' | S = s, A = a)$.
- γ: a discount factor, $\gamma \in [0, 1]$.
- $\mathcal{A}_i$: a set of possible actions for agent i.
- O_i: a set of observations for agent i.
- R_i: a reward function for the ith agent, $R_i(s, a)$.

At a given time t the agent senses the environment and receives a local observation $o_i : \mathcal{S} \rightarrow O_i$, based on the observation o_i and its stochastic policy $\pi_i : O_i \times \mathcal{A}_i \rightarrow [0, 1]$, the agent takes an action within the action-space $a_i \in \mathcal{A}_i$. Consequently, the agent transits to the next state s' which is determined based on the state transition probability function $P(s'|s, a) : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_N \rightarrow \mathcal{S}$ and receives a decentralized reward $r_i : \mathcal{S} \times \mathcal{A}_i \rightarrow \mathbb{R}$. The goal of each agent i is to optimally solve the POSG by deriving a probability distribution over actions in $\mathcal{A}$ at a given state, that maximizes its cumulative discounted sum of future rewards over an infinite time horizon and find the corresponding optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$.

An optimal policy maximizes the action-value function, i.e.,

$$\pi^*(s) = \arg \max_a Q^*(s, a) \qquad (1)$$

where,

$$Q^{\pi}(s, a) := \mathbb{E}_{\pi}[\sum_{k=0}^{\infty} \gamma^k R_k(s, a)|s_0 = s, a_0 = a]. \tag{2}$$

The optimal action-value function is determined by solving the Bellman equation,

$$Q^*(s, a) = \mathbb{E}\left[R(s, a) + \gamma \max_{a'} Q^*(s', a')|s_0 = s, a_0 = a\right] \tag{3}$$

3.1 Double Deep Q-Network

Deep Q-network (DQN) has been widely used in RL problems. DQN uses a deep neural network (NN) with weights $\mathbf{w}$ as the function approximator to estimate the state-action value function, i.e., $\tilde{Q}(.; \mathbf{w}) \cong Q(.)$. DDQN improves DQN by decomposing the max operation in the target into action selection and action evaluation, mitigating the over-estimation problem. The idea is to periodically sample data from a buffer and compute an estimate of the Bellman error or loss function, written as

$$\mathcal{L}(\mathbf{w}) = \mathbb{E}_{s,a,r,s' \sim \mathcal{RM}}[(Target - \tilde{Q}(s, a; \mathbf{w}))^2] \tag{4}$$

$$Target = R(s, a) + \gamma \tilde{Q}(s', \arg\max_{a'} \tilde{Q}(s', a'; \mathbf{w}); \hat{\mathbf{w}})) \tag{5}$$

The DDQN algorithm then performs mini-batch gradient descent steps as $\mathbf{w}_{i+1} = \mathbf{w}_i - \alpha_i \hat{\nabla}_{\mathbf{w}} \mathcal{L}(\mathbf{w})$, on the loss $\mathcal{L}$ to learn the approximation of the value function ($\tilde{Q}(.)$). The $\hat{\nabla}_{\mathbf{w}}$ operator denotes an estimate of the gradient at $\mathbf{w}_k$, $\mathbf{w}$ are the weights of the online network and $\hat{\mathbf{w}}$ are the weights of the target network which are updated at a lower frequency ($Target_{update}$) to stabilize training. The experience replay buffer (RM) is used to generate training samples (s, a, r, s'), which are randomly drawn to protect from correlated observations and non-stationary data distribution.

3.2 Driving Scenarios

Our objective is to investigate driving scenarios in which the lack of AV coordination hinders safety and efficiency. We also study adaptability among scenarios and driver behaviors. For this, we design a set of scenarios $\mathcal{F}$ with highway exiting and merging ramps as the main scenarios, as shown in Fig. 2, where a mission vehicle (in our

case an exiting/merging vehicle) attempts to accomplish its task in a mixed-traffic environment.

The exiting and merging scenarios are designed in such a way that coordination is necessary for safety. AVs must coordinate, and neither can achieve a safe and smooth traffic flow on its own, i.e., exiting/merging will not be feasible without the coordination of the other AVs. To facilitate safe exiting/merging while also responding to varied traffic scenarios, altruistic AVs must learn to account for the interests of all vehicles, coordinate, make compromises, and influence human behavior. In Fig. 2, for example, the AV1 has to compromise its own utility and reduce speed to guide the traffic of aggressive HVs, creating space for the exiting/merging vehicle, while the other AVs have to increase speed to create room for the mission vehicle. The exiting and merging scenarios are defined as f_e, $f_m \in \mathcal{F}$ correspondingly. We particularly chose those scenarios as a case of study because of their intrinsic similarity and the need for coordination, as the exiting/merging vehicle's utility contrast with that of the HV highway vehicles.

3.3 Social Value Orientation for AVs

In this section, we introduce *Social Value Orientation (SVO)* to formally investigate the social conflicts between humans and agents in diverse environments. It is critical to quantify an individual's social preference to understand whether they would cooperate or not in a particular scenario, such as opening a gap in our highway merging example. For that purpose, SVO is a commonly used concept in the social psychology literature that has lately been applied in robotics research [2]. In our context, SVO defines the degree of an agent's egoism or altruism toward others. Based on the value placed on the utility of others, an HV or an AV's behavior can range from egoistic to completely altruistic. We rely on AVs to guide traffic toward more socially advantageous outcomes since the SVO of HV is unknown. In formal terms, an AV's SVO angle ϕ determines how the AV balances its own reward against that of others [10, 17, 53]. In terms of rewards, an AV's total reward R_i is defined as:

$$R_i = r_i \cos \phi_i + r_i^- \sin \phi_i \qquad (6)$$

where r_i is the agent's individual utility, r_i^- is the total utility of other agents from the perspective of the ith agent which in general is a function $f(.)$ of their individual utilities,

$$r_i^- = f(r_j), \quad \text{where } j \neq i \qquad (7)$$

The SVO angle can varied from $\phi = 0$ (entirely selfish) to $\phi = \pi/2$ (entirely altruistic). Nonetheless, none of the limits are optimal, and a point in the middle, known as the optimal SVO angle ϕ^* gives the most socially favorable outcome.

Fig. 5 The SVO angle ϕ quantifies the level of altruism of an agent. In the figure the diameter of the circles, represents the size of the human population that holds the associated SVO [55]

SVO allows us to understand the behaviors that make possible the socially desirable outcomes in Fig. 2.

Autonomous agents must be aware of human drivers' social preferences as well as their desire to collaborate. Humans, on the other hand, are known to be diverse in SVO, and so their preferences are uncertain [54]. Figure 5 depicts a range of altruism across individuals with varying SVO. As a result of the wide range of altruistic behavior seen in humans, is not safe to rely on humans to guide the traffic, instead, we should rely on AV to guide the traffic toward more socially advantageous goals. Therefore, our objective is that the AVs learn to create alliances and influence HV behavior to improve the global utility of the group.

3.4 Autonomous Vehicles as Social Actors

AVs in a mixed environment will be social actors in the traffic road that will react to HVs and influence and adapt to their behaviors. The traffic environment is rich with habits and expectations, that determine driver behaviors. The vehicle's interaction will be influenced by the way AV decisions are perceived [2, 56]. For instance, some human drivers may be grateful if the AV stops for them but frustrated if it does not perform as expected. Also, they might behave aggressively if they're stuck behind an overly cautious AV, which reduces speed constantly. Another example is the case that when crossing a street while a vehicle is waiting, pedestrians move faster (a gesture of respect for the driver). On the other hand, will pedestrians speed up for an AV, or will they behave differently? If an AV is understood as a social actor, the HVs will learn the individual and social traits of AVs and behave accordingly in mutual interactions. This would fit with current preconceptions that

make assumptions about drivers based on the brand and type of vehicle they drive. Current AVs' driving is as conservative as possible to ensure safety. They will slow down in front of a crossing because they believe the other vehicle will want to go first, even though this is against the law. They wait for pedestrians when in doubt. It's not difficult to see how other agents and HVs could take advantage of and exploit these over-conservative behaviors. As AVs are going to be social actors in mixed autonomy traffic, the safety and reliability of AVs will be coupled with their social awareness and their ability to engage in complex social interactions. We consider risk awareness and social behavior as fundamental traits for decision-making.

Failure to identify social cues of selfishness or collaboration by an AV has ramifications for the general flow of the traffic network, as well as the safety of traffic participants. Current AVs ignore social signs and driver personality in favor of explicit communication or driver modeling. Because these methods can't handle complicated interactions, they tend to be conservative, restricting autonomy solutions to simple road interactions [2, 56]. The ability of AVs to drive in a socially obedient manner is critical for the safety of passengers and other vehicles because predictable behavior allows humans to comprehend and respond appropriately to the AV's actions.

3.5   Driving Behaviors

The problem of simulating varied behaviors may be defined as determining the appropriate range of parameters to produce heterogeneous behaviors within the simulator. Some works in social traffic psychology show that driving behavior falls between conservative and aggressive. Nevertheless, the specific definition is still under discussion and fluctuates across works [3]. The phrase "aggressive driving" refers to a wide range of unsafe driving practices, including running red lights and speeding. The root of aggressive driving has a variety of factors that aren't necessarily clear. Some are caused by hazardous road conditions, while others are caused by personal characteristics or mental states. Moreover, there is a correlation between aggressiveness and egoism, as egoistic drivers are less likely to yield and have a tendency to over-speeding and engage in unsafe actions. While there is a correlation between these concepts [12, 13], we distinguish aggressiveness from egoism in this study by describing individual traits and social preferences.

In this work, we discriminate between individual traits and social preferences because they result in different behaviors. We define altruism and egoism as social preferences; in that sense, an egoistic driver is a selfish driver who accounts for his personal utility irrespective of his aggression. We define conservatism and aggressiveness as individual traits, and we describe an aggressive driver as someone whose actions result in aggressive behavior. Individual traits such as aggressiveness are characterized by the outcomes of their actions, but social preferences such as egoism are distinguished by their social objectives and purposes. In this direction, an egoistic driver is a self-centered driver who lacks social motive, a driver who

believes he controls the road and disregards the other drivers. Egoist drivers frequently engage in violent actions, and while ego defensiveness is not the primary source of aggression, it is a major contributor to aggressive driving [12, 13]. Despite their similarities, the two groups have different origins and result in different behaviors. A driver, for example, could be egoistic and conservative. We may envision a driver who drives cautiously to protect himself (selfish motivation/preference) and, as a result, is conservative in his behavior (outcome of his actions).

Properly, we described social preferences (altruism or egoism) by the AV's SVO angular phase ϕ; and individual traits (conservativeness and aggressiveness) by the HV driver model parameters ($\mathcal{P}$) as described in Sect. 5.5. Based on the values of these parameters, a driver will behave conservative or aggressive. In the simulations, the AVs have no access to HVs' SVO, we consider the SVO of HVs to be undetermined as they cannot communicate that directly. Finally, we define a set of behaviors $\mathcal{B}$, i.e, aggressive, moderate and conservative, $b_a, b_m, b_c \in \mathcal{B}$ based on the parameters ($\mathcal{P}$) obtained in Sect. 5.5.

4 Problem Formulation

We investigate the safety and robustness of the scenarios described in Fig. 2, an exiting/merging vehicle, which can be either HV or AV. This configuration contains a group of AVs that hold the same SVO, as well as a group of HVs which are heterogeneous in their SVO, making it unclear whether they are allies or opponents. Formally, the road is shared by a set of HVs $h_k \in \mathcal{H}$, with an undetermined SVO ϕ_k and heterogeneous behaviors $b_k \in \mathcal{B}$; a set of AVs $i_i \in \mathcal{I}$, that are connected together using V2V communication, controlled by a decentralized policy and sharing the same SVO, and a *mission vehicle*, $M \in \mathcal{I} \cup \mathcal{H}$ that is aiming to accomplish its mission (highway exiting/merging) and it can be either AV or HV. We focus on the multi-agent maneuver-level decision-making problem for AVs in mixed-autonomy environments and study the following problems: how AVs can learn in a mixed-autonomy environment optimal cooperative policies $\pi^*(s)$ that are robust to different scenarios $f \in \mathcal{F}$ and behaviors $b \in \mathcal{B}$ while ensuring safety on the decision-making, and how sensitive is the performance of the altruistic AVs to the HVs' behaviors.

As AVs are connected, we assume that they receive an accurate local observation of the environment $\tilde{\mathbf{o}}_i \in \widetilde{O}_i$, sensing all the vehicles within their perception range, i.e, a subgroup of HVs $\widetilde{\mathcal{H}} \subset \mathcal{H}$ and a subgroup of AVs $\widetilde{\mathcal{I}} \subset \mathcal{I}$. Nevertheless, AVs are unable to share their actions or rewards, and they take individual actions from a set of high-level actions $a_i \in \mathcal{A}_i (|\mathcal{A}_i| = 5)$. The goal of this work is to train social-aware AVs that learn how to drive in a mixed-autonomy scenario in a robust, efficient, and safe manner. We are interested in how to obtain a utility function that enables AVs to handle competitive driving scenarios (such as those in Fig. 2) and leads them into socially-desirable decisions that improve traffic efficiency, safety, and robustness.

5 Safe and Robust Social Driving

In this section, we present the safe and robust MARL approach. Our approach uses a general decentralized reward function that optimizes for social utility and induces altruism in the AVs; the general reward function accounts for any anticipated vehicle's mission, allowing it to be applied to a variety of environments; and collisions are reduced by the safety prioritizer. What we define as "driving" is the outcome of decades of human learning from experience. Consequently, we take the same approach and train AVs that learn from experience and define the optimization problem as the eventual desirable social outcome with adaptability, expecting AVs to learn how to drive safely during the process. We carefully design a decentralized general reward function, a suitable architecture, and a safety prioritizer to promote the desired safe altruistic behavior in AVs' decision-making process. The overview of our approach as presented in Figs. 4 and 2 helps us to create intuition on these points, by introducing driving scenarios in which altruistic AVs lead to socially advantageous results while adapting to different traffic scenarios.

Action Space The goal of this research is to look at inter-agent and agent-human interactions, as well as behavioral elements of mixed-autonomy driving. Thus, we choose a more abstract level and define the action-space as a set of discrete meta-actions $a_i \in \mathcal{A}_i$. In particular, we select a set of five high-level actions a_i as,

$$a_i \in \mathcal{A}_i = \begin{bmatrix} \texttt{LaneLeft} \\ \texttt{Idle} \\ \texttt{LaneRight} \\ \texttt{Accelerate} \\ \texttt{Decelerate} \end{bmatrix} \tag{8}$$

These meta-actions are then converted into trajectories and low-level control signals, which ultimately control the vehicle's movement.

Observation Space We use a *multi-channel VelocityMap* observation (o_i) that embeds the relative speed of the vehicle with respect to the ego vehicle in pixel values [17]. We represent the information in multiple semantic channels that embed: (1) an attention map to highlight the position of the ego vehicle, (2) the HVs, (3) the AVs, (4) the mission vehicle, and (5) the road layout. Figure 6 illustrates an example of this multi-channel representation. In order to map the relative speed of the vehicles into pixels, we use a clipped logarithmic function, which improves dynamic range and yields better results than a linear map, i.e.,

$$Z_j = 1 - \beta \log(\alpha |v_j^{(l)}|) \mathbb{1}(|v_j^{(l)}| - v_0) \tag{9}$$

where Z_j is the pixel value of the jth vehicle in the state representation, $v^{(l)}$ is its relative Frenet longitudinal speed from the kth vehicle's point-of-view, i.e.,
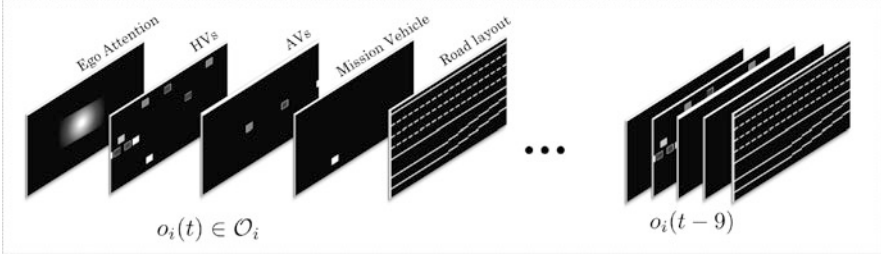
Fig. 6 Multi-channel VelocityMap state representation embeds the speed of the vehicle in pixel values

$\dot{l}_j - \dot{l}_k$, v_0 is speed threshold, α and β are dimensionless coefficients, and $\mathbb{1}(.)$ is the Heaviside step function. Such non-linear mapping gives more importance to neighboring vehicles with smaller $|v^{(l)}|$ and almost disregards the ones that are moving either much faster or much slower than the ego vehicle. As temporal information is necessary for safe decision-making, we use a history of successive VelocityMaps observations to create the input state to the Q-network.

5.1 Distinguishing Sympathy from Cooperation

In our mixed-autonomy problem, we divide inter-agent relations into interactions between autonomous agents (AV-AV interactions) and interactions between autonomous agents and human drivers (HV-AV interactions). By decoupling the two, we can analyze the interactions between human drivers with unclear SVO and our autonomous agents in a methodical way. In that sense, we define *sympathy* as the autonomous agent's altruism toward a human, and *cooperation* as the altruistic behavior among autonomous agents. The fact that the components of altruism differ in nature is our reasoning for separating them. Sympathy, for example, may not be reciprocated since agents differ in their SVO, whereas cooperation among autonomous entities is fundamentally homogeneous if they share the same SVO. Following this concept, we can rewrite the AV reward in Eq. (6) as,

$$
\begin{aligned}
R_i &= r_i \cos \phi_i + (\sin \theta_i R_i^{\text{AV}} + \cos \theta_i R_i^{\text{HV}}) \sin \phi_i \\
&= \underbrace{r_i \cos \phi_i}_{\text{egoistic term}} + \\
&\quad \underbrace{\sin \theta_i \sin \phi_i R_i^{\text{AV}}}_{\text{cooperation term}} + \underbrace{\cos \theta_i \sin \phi_i R_i^{\text{HV}}}_{\text{sympathy term}}
\end{aligned}
\tag{10}
$$

where θ is the sympathy angular phase determining the cooperation-to-sympathy ratio. Parameters R_i^{AV} and R_i^{HV} denote the total utility of other AVs and HVs, respectively, as perceived from the ith agent's perspective. We expand on this topic in Sect. 5.2 where we introduce the distributed reward structure.

5.2 Decentralized Social Reward

The AVs are trained using the partial local observations and the decentralized reward function, and we anticipate them to learn how to drive in a variety of settings while taking into consideration the individual diver's missions. As a result, we create a well-engineered general reward function that considers social utility, traffic metrics, and individual diver's missions. Following the definition of sympathy and cooperation in equation (10) we decompose the decentralized reward received by agent $I_i \in \mathcal{I}$ as,

$$
\begin{aligned}
R_i(s, a) &= R^{\text{ego}} + R^{\text{social}} \\
R^{\text{ego}} &= \cos \phi_i \, r_i(s, a) \\
R^{\text{social}} &= R^{\text{coop}} + R^{\text{symp}} \\
R^{\text{coop}} &= \sin \theta_i \sin \phi_i \Big[\sum_j r_{i,j}^{AV}(s, a) + \sum_j r_{i,j}^M(s, a) \Big] \\
R^{\text{symp}} &= \cos \theta_i \sin \phi_i \Big[\sum_k r_{i,k}^{HV}(s, a) + \sum_k r_{i,k}^M(s, a) \Big]
\end{aligned}
\tag{11}
$$

in which R^{ego}, R^{social} represents the egoistic and social reward, $i \in \mathcal{I}$, $j \in (\widetilde{\mathcal{I}} \setminus \{I_i\})$, $k \in \widetilde{\mathcal{H}}$. The term r_i represents the ego vehicle's reward obtained from traffic metrics and the angle ϕ allows to adjust the level of egoism or altruism. R^{coop} is the cooperation term (the altruistic behavior among AVs, i.e, AV's altruism toward others AVs) and R^{symp} is the sympathy term (AV's altruism toward HVs). The sympathy reward term, $r_{i,k}^{HV}$ considers the individual reward of the HVs, while the cooperation reward term, $r_{i,j}^{AV}$ considers the individual reward of the other AVs, and are defined as

$$
r_{i,k}^{HV} = \frac{\mathcal{W}_k}{d_{i,k}^{\lambda}} \sum_m \omega_m x_m \quad r_{i,j}^{AV} = \frac{\mathcal{W}_j}{d_{i,j}^{\lambda}} \sum_m \omega_m x_m
\tag{12}
$$

in which $d_{i,k}/d_{i,j}$ represents the distance between the agent and the corresponding HV/AV, λ is a dimensionless coefficient, $\mathcal{W}_k$ is a weight value for individual vehicle's importance, m is the set of traffic metrics that have been considered in the vehicle's utilities (speed, crashes, etc.), in which x_m is the m metric normalized

value and w_m is the weight associated to that metric. The term r^M accounts for the reward of the vehicle's mission. A mission is defined as any desired specific outcome for a particular vehicle, as merging, exiting, etc.

$$r_{i,j}^M = \begin{cases} \frac{w_j}{(d_{i,j})^\mu}, & \text{if } g(j) \\ 0, & \text{o.w.} \end{cases} \qquad r_{i,k}^M = \begin{cases} \frac{w_k}{(d_{i,k})^\mu}, & \text{if } g(k) \\ 0, & \text{o.w.} \end{cases} \tag{13}$$

The function $g(v)$ is an independent function to evaluate the mission; $g(v)$ returns true if the vehicle v has a mission defined and the mission has been accomplished in the recent time window. μ is a dimensionless coefficient, w_j/w_k are weights for an individual vehicle's mission (importance of the mission). This allows defining a general reward independent of the driving scenario and mission goals for different vehicles. In the experiments, a **HV** can be assigned a merging mission or a highway exiting mission, as referred to in Fig. 2.

5.3 Deep MARL Architecture for Social Driving

As shown in Fig. 8, we leverage a 3D Convolutional Neural Network (CNN) with a safety prioritizer for our MARL architecture. To account for the temporal information, the 3D CNN operates as a feature extractor and leverages a history of VelocityMap observations. The network receives a stack of 10 VelocityMap observations, i.e., a $10 \times (4 \times 512 \times 64)$ tensor that captures the latest 10 time-steps episodes. To mitigate the non-stationarity issue in MARL, agents are trained in a semi-sequential manner, as illustrated in Fig. 7. The agents are trained independently for $N_{iterations}$ iterations while freezing the policies of the remaining AVs, $\mathbf{w}^-$. Subsequently, the other agents' policies are updated with the new policy, $\mathbf{w}^+$.

To improve safety we train our agents using a safety prioritizer that, in the cases where the action selected by the agent policy is unsafe, selects a safe action and stores the unsafe action (a_t) and the related state in the RM with a suitable penalty on the reward (r_{unsafe}) for the unsafe state-action pair. The safety prioritizer reduces episode resets due to imminent collisions improving sample efficiency. The unsafe state-action pairs are not removed so the agent can also learn from unsafe experiences. The experience ($\psi(s_t), a_t, r_{unsafe}, \emptyset$) is stored in RM with a terminal next state $\emptyset$, the target for this unsafe pair (s_t, a_t) is $Target(s_t, a_t)^{DDQN} = r_{unsafe}$. The details of the safety prioritizer are given in the next Sect. 5.4.

The proposed deep MARL architecture is described in Algorithm 1. As part of the implementation, we start the learning process after the replay buffer has been filled with a sufficient number of sample simulations. Furthermore, we update the experience replay buffer to adjust for the extremely skewed training data [17]. Balancing skewed data is a frequent practice in machine learning, and it was effective in our MARL problem.
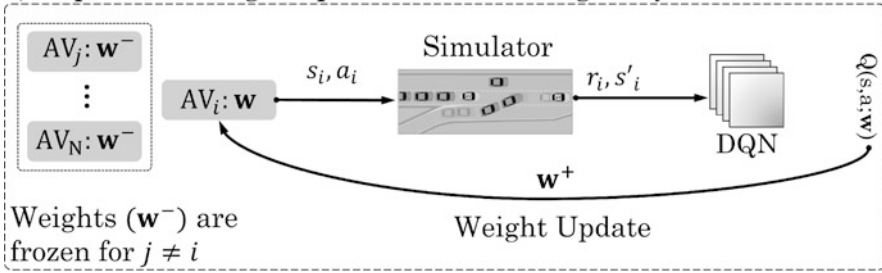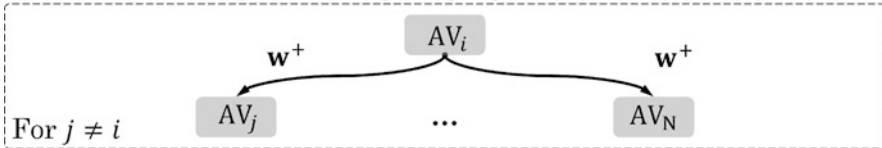
Fig. 7 The multi-agent training and policy dissemination process

5.4 Safety Prioritizer

We include a safety prioritizer to the MARL algorithm that penalizes and reduce imminent crashes. This helps the agent to increase sample efficiency during training and avoid collisions when in deployment. If the agent comes into an unexpected situation and decides to perform a risky action, that action will be prevented. The safety prioritizer enhances simulation results and is crucial in real-world scenarios. The safety prioritizer included Algorithms 2 and 3.

Algorithm 2 During action selection of the agent I_i, once an action a_t is chosen, the safety prioritizer checks if the action is safe by computing a safety score for N_{steps} of planning. We utilize the time-to-collision (ttc) as a safety score. If $safety_{score} < safe_{th}$ the action is unsafe and we need to select a safe action. The selection of a safe action is presented in Algorithm 3.

Algorithm 3 The safe action selection is different in training and testing. During training, to encourage exploration, we remove the unsafe actions and keep the random action selection following the current exploration policy on the remaining actions. During testing, we follow the greedy policy in the subset of safe actions $a_t = \max_{a' \in \tilde{\mathcal{A}}_{safe}} Q(\psi(s_t), a'; \mathbf{w})$. It should be noted that the algorithm does not choose the safest of all possible actions, as that action may lead to particularly conservative behaviors that can compromise traffic efficiency; we instead remove the imminent unsafe actions and follow the priority given by the learned altruistic policy. If it happens that all possible actions are unsafe, we return the action $a_t \in \mathcal{A}$ with the highest safety score. In that way during training the constrained exploration will keep the agent from taking unsafe actions which will lead to efficient sampling

---

Algorithm 1 Safety Prioritized Multi-agent DDQN

Initialize *experience replay buffer RM*.
Initialize $\tilde{Q}(.; \mathbf{w}^-)$ with random weights $\mathbf{w}^- = \mathbf{w}_{ini}$
Initialize target network $\tilde{Q}(.; \hat{\mathbf{w}})$ with weights $\hat{\mathbf{w}} = \mathbf{w}^-$
Pre-store experience of first's 50 episodes in *RM*
for e = 50 to N_{episode} **do**
   Initialize $s_1 = \{\tilde{\mathbf{o}}_1\}$ and compute $\psi_1 = \psi(s_1)$
   for t = 1 to T **do**
      for I_i in $\mathcal{I}$ **do**
         Freeze $\mathbf{w}^-$ for all I_j, $j \neq i$
         for $m = 1$ to $N_{iterations}$ **do**
            With probability ϵ select a random action a_t,
            otherwise select $a_t = \max_{a' \in A} Q(\psi(s_t), a'; \mathbf{w}^+)$
            if a_t is unsafe (Algorithm 2) **then**
               Store $(\psi_t, a_t, r_{unsafe}, \emptyset)$ in *RM*
               a_t = Compute a safe action (Algorithm 3)
            end if
            Execute safe action a_t, and observe $r_t, \tilde{\mathbf{o}}_t$
            Set $s_{t+1} = \{s_t, \tilde{\mathbf{o}}_{t+1}\}$ and $\psi_{t+1} = \psi(s_{t+1})$
            Store experience $(\psi_t, a_t, r_t, \psi_{t+1})$ in *RM*
            Sample a mini-batch of size M from *RM*
            Compute $\mathcal{L}(\mathbf{w}^+)$
            Performs gradient descent
            $\mathbf{w}_{k+1}^+ \leftarrow \mathbf{w}_k^+ - \alpha \hat{\nabla}_{\mathbf{w}} \mathcal{L}(\mathbf{w}^+)$
         end for
         $\mathbf{w}^- = \mathbf{w}^+$ for all $I_i \in \mathcal{I}$
      end for
      Every $Target_{update}$ reset $\hat{\mathbf{w}} \leftarrow \mathbf{w}^-$
   end for
end for

Algorithm 2 Safety score

Simulate I_j taking the action a_t
for v in $(\mathcal{I} \cup \widetilde{\mathcal{V}}) \setminus \{I_i\}$ **do**
   Compute safety score of I_i, v for N_{steps} planning
   if $safe_{score} < safe_{th}$ **then**
      Return unsafe
   end if
end for
Return safe

---

and more stable learning; and during testing, the decision-making is based on the prosocial learned policy with minimum intervention from the safety prioritizer, achieving higher traveled distance while avoiding collisions (Fig. 8).

Algorithm 3 Safe action

Initialize $\widetilde{\mathcal{A}} = \mathcal{A}$
while $\widetilde{\mathcal{A}}$ is not empty **do**
    if during training **then**
        Select a_t following the exploration policy on set $\widetilde{\mathcal{A}}$
    else if during test **then**
        Select $a_t = \max_{a' \in \widetilde{\mathcal{A}}_{safe}} Q(\psi(s_t), a'; \mathbf{w})$
    end if
    if a_t is safe (Algorithm 2) **then**
        Return a_t
    end if
end while
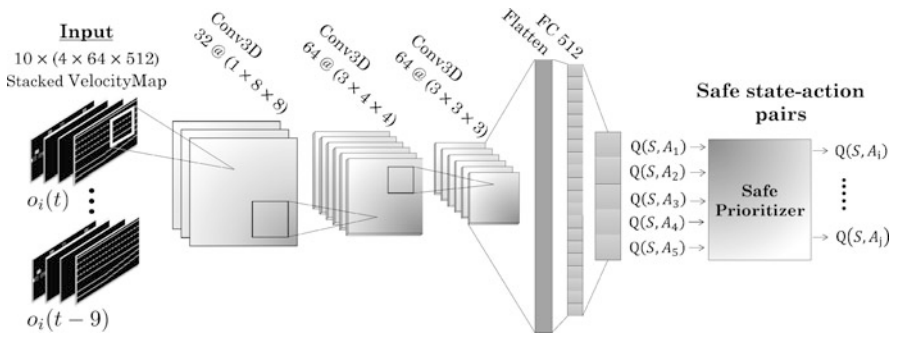Return a_t with highest safe score in $\mathcal{A}$

Fig. 8 Deep MARL architecture with the safety prioritizer

5.5 Modeling Driver Behaviors

We model the longitudinal movements of HVs using the *Intelligent Driver Model* (IDM) [57], while the lateral actions of HVs are based on the MOBIL model [58]. The MOBIL model considers two main criteria,

The safety criterion ensures that after the lane change, the deceleration of the new follower a_n in the target lane does not exceed a safe limit, i.e, $a_n > -b_{\text{safe}}$.

The incentive criterion determines the advantage of HV after the lane change, quantified by the total acceleration gain, given by

$$a'_{ego} - a_{ego} + \sin\phi_{ego}\Big((a'_n - a_n) + (a'_o - a_o)\Big) > \Delta a_{th} \tag{14}$$

where a_o, a_n and a_{ego} represent the acceleration of the original follower in the current lane, the new follower in the target lane and the ego HV, correspondingly, and a'_o, a'_n, and a'_{ego} are the equivalent accelerations considering that the ego HV has changed the lane, $\sin\phi_{ego}$ is the politeness factor. Finally, the lane change is performed if the safety and incentive criteria are mutually satisfied.

The IDM Model determines the longitudinal acceleration of a HV $\dot{v}_k$ as follows,

$$\dot{v}_k = a_{max}\left[1 - \left(\frac{v_k}{v_k^0}\right)^\delta - \left(\frac{d^*(v_k, \Delta v_k)}{d_k}\right)^2\right] \tag{15}$$

in which v_k, d_k, δ, Δv_k, v_0^k denote the speed, the actual gap, the acceleration exponent, the approach rate, and the desired speed of the kth HV, respectively.

The desired minimum gap of the kth HV is given by,

$$d^*(v_k, \Delta v_k) = d_k^0 + v_k T_k^0 + \frac{v_k \Delta v_k}{(2\sqrt{a_{max} \cdot a_{des}})} \tag{16}$$

where T_k^0, d_k^0, a_{max}, and a_{des} are the safe time gap, the minimum distance, the comfortable maximum acceleration, and deceleration, correspondingly.

The typical parameters for the MOBIL model are $\sin\phi_{ego} = 0.5$, $\Delta a_{th} = 0.1\frac{m}{s^2}$ and $b_{safe} = 4\frac{m}{s^2}$. Table 1 shows typically used parameters of the IDM model [57].

Heterogeneous Driver Behaviors Although those parameters are typically used for IDM and MOBIL models, they simulate just one behavior. In order to generate diverse behaviors $\mathcal{B}$, we frame the task of simulating diverse behaviors as the problem of obtaining the appropriate range of parameters ($\mathcal{P}$) that can generate those behaviors. To achieve that, we leverage a behavior classifier and iteratively simulate the parameters and classify the behaviors, mapping parameters to behaviors. To classify the behaviors we represent traffic using a traffic-graph at each time step t, $\mathcal{G}_t$, with a set of edges $\mathcal{E}(t)$ and a set of vertices $\mathcal{V}(t)$ as functions of time, i.e, the positions of vehicles ($\mathcal{H} \cup \widetilde{\mathcal{I}}$) represent the vertices. The adjacency matrix A_t is given by $A(k, m) = d(v_k, v_m), k \neq m$, in which $d(v_k, v_m)$ is the shortest travel distance between vertices k to m. Then we use centrality functions [34] to classify the behavior (level of aggressiveness) resulting from $\mathcal{P}$, and then use those simulation parameters $\mathcal{P}$ to model behaviors within the simulator with varying levels of aggressiveness. The centrality functions are defined as,

Closeness Centrality the discrete closeness centrality of the kth vehicle at time t is defined as,

$$C_C^k[t] = \frac{N - 1}{\sum_{v_m \in \mathcal{V}(t) \setminus \{v_k\}} d_t(v_k, v_m)}, \tag{17}$$

Table 1 Common used parameters for the IDM model

Parameter	v^0	T^0	a_{max}	a_{des}	δ	d^0
Value	30 m/s	1.5 s	1 m/s^2	1.5 m/s^2	4	2 m

The more central the vehicle is located, the higher $C_C^k[t]$ and the closer it is to all other vehicles.

Degree Centrality the discrete degree centrality of the kth vehicle at time t is defined as,

$$C_D^k[t] = \left| \{v_m \in \mathcal{N}_k(t)\} \right| + C_D^k[t-1]$$
$$\text{such that } (v_k, v_m) \notin \mathcal{E}(\tau), \tau = 0, \ldots, t-1 \tag{18}$$

in which $\mathcal{N}_k(t) = \{v_m \in \mathcal{V}(t), A_t(k,m) \neq 0, v_m \leq v_k\}$ represents the set of vehicles in the proximity of the kth vehicle, given that $v_m \leq v_k$; and v_m, v_k denote the velocities of the mth and kth vehicles, $A_t(k,m)$ is the adjacency matrix. The more new vehicles seen by vehicle k that meet this condition, the higher $C_D^k[t]$.

With the centrality functions, we can measure the Style Likelihood Estimate (SLE) for different driver styles [34]. We consider two SLE measures. The SLE of overtaking and sudden lane changes (SLE_l) and the SLE of overspeeding (SLE_o). The SLE_l and SLE_o can be computed by measuring the first derivative of the centrality functions as,

$$\text{SLE}_l(t) = \left| \frac{\partial C_C(t)}{\partial t} \right| \quad \text{SLE}_o(t) = \left| \frac{\partial C_D(t)}{\partial t} \right| \tag{19}$$

The maximum likelihood $\text{SLE}_{\max}$ is calculated as $\text{SLE}_{\max} = \max_{t \in \Delta t} \text{SLE}(t)$.

Using those functions, we can approximately quantify and classify driver behaviors in our simulation. The intuition behind that is that an aggressive driver may frequently overspeed or perform sudden lane changes; while overspeeding the $C_D(t)$ monotonically increases (higher $\text{SLE}_o(t)$) and during sudden lane changes the slope and the extrema of $C_C(t)$ changes values. Thus higher values of $\text{SLE}_{\max}$ are related to increased levels of aggressiveness. Conversely, conservative drivers are not inclined toward those aggressive maneuvers, and the degree of centrality will be relatively flat, thus $\text{SLE}_o(t) \approx 0$ for conservative drivers.

We use these metrics as approximations of the driver's level of aggressiveness. In order to compute the suitable values for our simulation, we iteratively simulate the parameters from IDM and MOBIL models, and for each set of parameters, we quantify the resulting behavior in the simulation (using those metrics). A mapping of the parameters $\mathcal{P}$ to behaviors (quantified in the simulation for those parameters). The estimated simulation parameters that simulate conservative, moderate and aggressive behavior in our scenarios are presented in Table 2.

The desired velocity v^0 is set to $30 m/s$ and the acceleration exponent $\delta = 4$.

Table 2 Estimated simulation parameters for conservative, moderate, and aggressive behaviors

Model	Parameter	Aggressive	Moderate	Conservative
MOBIL	$\sin\phi_e$	0	0.3	1
	Δa_{th}	$0\,\text{m/s}^2$	$0.1\,\text{m/s}^2$	$0.4\,\text{m/s}^2$
	b_{safe}	$12.0\,\text{m/s}^2$	$6.0\,\text{m/s}^2$	$2.0\,\text{m/s}^2$
IDM	T^0	$0.5\,\text{s}$	$1\,\text{s}$	$3\,\text{s}$
	d^0	$1\,\text{m}$	$2\,\text{m}$	$6.0\,\text{m}$
	$\text{acc}_{\max}$	$7.0\,\text{m/s}^2$	$3.0\,\text{m/s}^2$	$1.0\,\text{m/s}^2$
	acc_{des}	$12.0\,\text{m/s}^2$	$7.0\,\text{m/s}^2$	$2.0\,\text{m/s}^2$

Table 3 Computation time for each agent

Computing platform	Online forward pass time
NVIDIA Tesla V100 GPU	3.7 ms
OnLogic Karbo 700 x2	65.2 ms
NVIDIA Jetson AGX Xavier GPU	32.9 ms
NVIDIA Jetson TX2 GPU	112.5 ms

5.6   Implementation and Computational Details

We customize the OpenAI Gym environment in [59] to suit our particular driving situation and MARL problem. We design a merging ramp and exiting highway scenario for our simulation running in python and used Pytorch for the implementation of our safety prioritized MARL DDQN algorithm. Our implementation on average uses 3.1GB of memory for 4 agents and 18 HVs using a GPU NVIDIA Tesla V100. The training process is repeated several times to ensure convergence of the experiments to a similar policy. The network is trained for $N_{episodes} = 10{,}000$ taking on average 8 h. While each round of 10, 000 training episodes in the Tesla V100 GPU takes around 8 h, a full forward pass during deployment for 4 simulated agents takes 15 ms (approximately 4 ms per agent).

In a real AV platform, each agent will receive a local observation of the environment that will be used by our algorithm to compute the safe optimal action based on the trained Q-network. The decision-making will take place on each AV's onboard computer; therefore, to verify the feasibility of the real-time operation of our decentralized algorithm we tested a forward pass of the Q-network during deployment in multiple hardware platforms. The results for the different platforms are presented in Table 3, for instance, an online forward pass of the network in the deployment phase using commodity GPU hardware, i.e, an NVIDIA Jetson AGX platform will be around 32.9 ms for each agent. We utilize 3200 GPU hours for all our simulation experiments. Table 4 lists our simulation and training hyper-parameters.

Table 4 List of hyper-parameters

Parameter	Value	Parameter	Value
N_{episode}	10,000	ϵ decay	Linear
RM buffer size	8000	Initial exploration ϵ_0	1.0
Batch size	32	Final exploration	0.05
Learning rate α_0	0.0005	Optimizer	ADAM
$Target_{update}$	300	Discount factor γ	0.95
$|\mathcal{H}|$	18	$|\mathcal{I}|$	4

6 Experiments and Results

6.1 Manipulated Variables

We study how the $safe_{th}$, the *level of aggressiveness*, the *traffic scenarios* (f_j) and the *HVs' behaviors* (b_k) impact the performance of AVs. We consider the case in which the mission vehicle (exiting/merging) in Fig. 2 is *human-driven*, $M \in \mathcal{H}$, and define the following terms:

- AV_S. Social AV ($\phi_i = \phi^*$) that act *altruistically* in the presence of diverse HVs behaviors $b \in \mathcal{B}$.
- AV_E. Egoistic AV ($\phi_i = 0$) that act *egoistically* in the presence of diverse HVs behaviors $b \in \mathcal{B}$.

with ϕ^* to be the optimal SVO angle tuned to reach the optimal level of altruism as in [17].

6.2 Performance Metrics

The performance of our system is measured based on safety, efficiency, altruistic performance gain (PG), and adaptation error A_{error}. To measure safety, we compute the percentage of episodes that encountered a crash ($C(\%)$). For efficiency, the average traveled distance ($DT(m)$) of the vehicles and the number of missions accomplished by the mission vehicle is used. The altruistic performance gain is measured by computing the difference in the safety/efficiency performance of AV_E and AV_S, as

$$PG_{safety}(\%) = \frac{(AV_E)_{C(\%)} - (AV_S)_{C(\%)}}{N_{Episodes}} \qquad (20)$$

$$PG_{efficiency}(\%) = \frac{(AV_S)_{DT(m)} - (AV_E)_{DT(m)}}{(AV_E)_{DT(m)}} \qquad (21)$$

Finally, the adaptation error is a weighted sum function of the safety $(C(\%))$ and efficiency $(DT(m))$ performance of the AV_S when trained and tested in different scenarios/behaviors. Defined as

$$A_{error}(\%) = w_s \times (C(\%)) + w_e \times 100(1 - \frac{DT}{DT_{max}}) \qquad (22)$$

such that an adaptation between different situations that result in 0% crash and $DT = DT_{max}$ will have $A_{error} = 0\%$.

6.3 Hypotheses

In this section we examine the following hypotheses

- **H1.** *In a mixed-autonomy scenario, the higher the level of aggressiveness, the bigger the impact of cooperation. We expect a higher performance gain (PG) when altruistic AVs face more aggressive environments.*
- **H2.** *Altruistic AVs agents using the decentralized framework can adapt to different driver behaviors and traffic scenarios without compromising the overall traffic metrics. However, the higher the similarity of testing scenarios to the ones seen during training (($f_{test}, b_{test}) \approx (f_{train}, b_{train})$), the lowest adaption error (A_{error}).*
- **H3.** *We anticipate an improvement in both safety and efficiency with the addition of the safety prioritizer. In the absence of a safety prioritizer ($safe_{th} = 0$) we expect that AVs will cause more crashes.*

6.4 Analysis and Results

Based on the hypotheses, we explore their correctness through the experiments in this section.

6.4.1 Sensitivity Analyses

To study the hypothesis **H1** we investigate the effect of HV behaviors on the altruistic AV agents. We focus on scenarios with a HV mission vehicle, with safe AVs that act *altruistically* (AV_S) or *egoistic* (AV_E), in environments with increasing levels of HVs aggressiveness. Figure 9 illustrates the altruistic performance gain for increasing levels of HVs' aggressiveness for 2 AVs (left) and 4 AVs (right). It demonstrates that the more aggressive the HVs are, the higher the impact of cooperation and thus confirms the **H1**. This is also observed in Fig. 10 where the level of aggressiveness is decomposed into lateral and longitudinal aggressiveness.
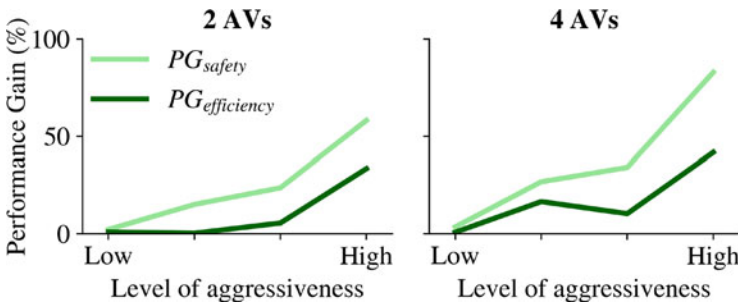
Fig. 9 Sensitivity analyses measured by altruistic performance gains (PGs) of AVs show that the more aggressive the HVs are, the more the impact/gain of cooperation
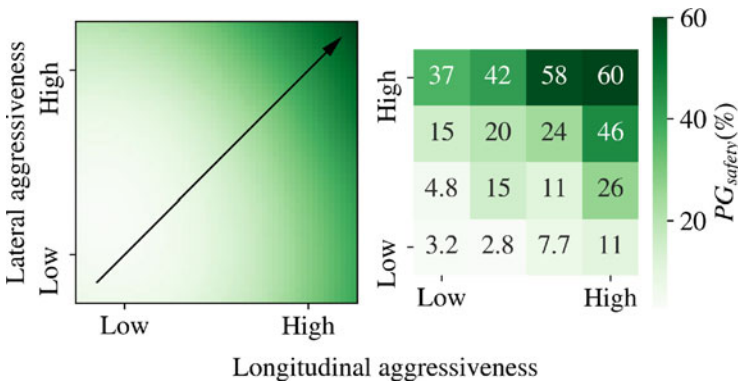

Fig. 10 Both lateral and longitudinal sensitivity analyses indicate an increase in altruistic performance gain (PG)

Lateral and longitudinal aggressiveness is varied by changing the MOBIL and IDM parameters (Table 2) from aggressive to conservative. Figure 10 shows that the altruistic gain increases in both directions, but is more pronounced in the longitudinal direction. That is probably due to the simulated scenarios having more longitudinal maneuvers.

6.4.2 Domain Adaptation

Following the sensitivity analysis, we investigate the domain adaptation of the AVs to validate the **H2**. Figures 11, 12 and 13 show how the altruistic AVs learn to adapt to different scenarios and behaviors by different performance metrics, i.e, crashed (**a**), distance traveled (**b**) and adaptation error (**c**). For the experiments, AV_S are trained in different scenarios $f_i \in \mathcal{F}$ in the presence of HVs with different behaviors $b_k \in \mathcal{B}$ and tested in other scenarios $f_j \in \mathcal{F}$ and behaviors $b_l \in \mathcal{B}$. In our experiments, we consider two case study scenarios $f_e, f_m \in \mathcal{F}$ (exiting/merging)
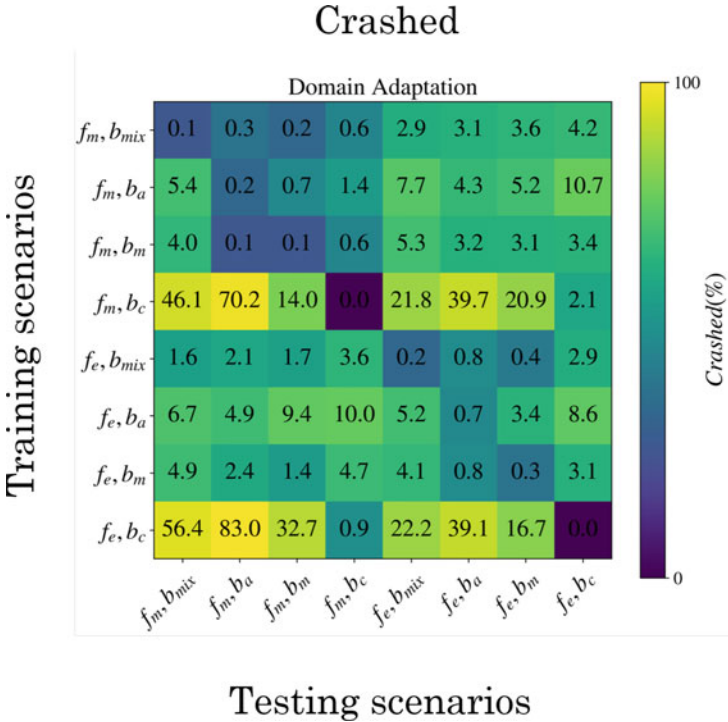
Crashed

Fig. 11 The domain adaptation matrix with crash percentage ($C(\%)$) between different traffic scenarios and behaviors. The lower $C(\%)$ the most suitable the adaptability in terms of safety (measured by $C(\%)$) between those domains. AV_S are trained (rows of the matrix) in different scenarios $f_i \in \mathcal{F}$ in the presence of HVs with different behaviors $b_k \in \mathcal{B}$ and tested (columns of the matrix) in other scenarios $f_j \in \mathcal{F}$ and behaviors $b_l \in \mathcal{B}$. Each pair (f_i, b_k) is a combination of scenario and behavior

in environments with three different HVs behaviors $b_a, b_m, b_c \in \mathcal{B}$ (aggressive, moderate, conservative) see Table 2; and a mixed behavior environment, in which HVs are created randomly and their behaviors are selected based on a uniform distribution over the behaviors in $\mathcal{B}$, given equal probability to the defined behaviors. In total, we have eight combinations of scenarios and behaviors, namely: (f_m, b_{mix}), $(f_m, b_a), (f_m, b_m), (f_m, b_c), (f_e, b_{mix}), (f_e, b_a), (f_e, b_m), (f_e, b_c)$.

The results are presented in Fig. 13 as an adaptation matrix, showing the A_{error} for different domains, the A_{error} is in percentage (%) and color-map in logarithmic scale to increase the perceived dynamic range for visualization. In our analyses, the weights used for $A_{error}(\%)$ are $w_s = \frac{2}{3}$ and $w_e = \frac{1}{3}$, which weighs the safety performance higher. DT_{max} is computed based on the maximum distance for each situation. Additionally, Figs. 11 and 12 illustrate how the AVs adapt in terms of safety (measured by $C(\%)$) and efficiency (measured by $DT(m)$), separately.

The matrix shows the best performances in its diagonal; where agents are trained and tested in the same environment $((f_i, b_k); (f_j, b_l)$ with $i = j$ and $k = l)$; due to the fact that agents experience similar situations during testing as they do during
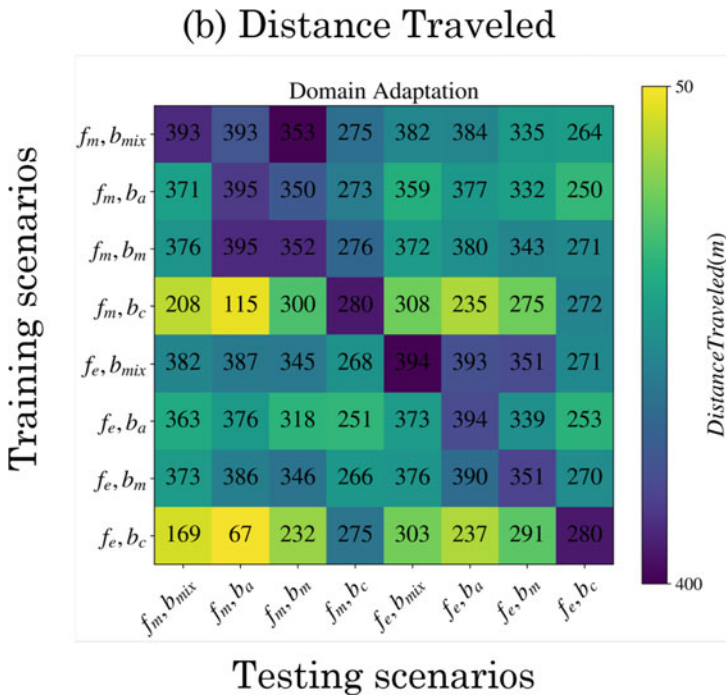
Fig. 12 The domain adaptation matrix with distance traveled ($DT(m)$). Illustrating how the AVs adapt to other situations in terms of efficiency (measured by $DT(m)$)

training. The vehicles trained in the merging environment can perform the exiting mission for different behaviors, and vice-versa. Interestingly, the performance of AVs trained in a conservative environment (b_c) is poor when tested in an aggressive environment (b_a). We believe that the reason is that in conservative environments, the HVs yield the mission vehicle, and the AVs learn to rely on HVs to guide the traffic. This learned policy is valid in a conservative environment where one can expect the HVs to always create a safe space for the mission vehicle. However, the same is not valid in more aggressive environments, in which AVs have to guide the traffic to avoid dangerous situations. As a result, the performance of vehicles trained in a conservative environment and tested in an aggressive one is the worse.

On the other hand, an adequate performance adaptation (lower A_{error}) is obtained when agents are trained in the presence of all moderate HVs (b_m) or a mixed behavior environment (b_{mix}), in which AVs face situations where the HVs yield, but also situations that require learning how to guide the traffic to optimize for the social utility. The results from the domain adaptation matrix indicate that a moderate or mixed environment is the most suitable for training robust AVs and show the adaptability of AVs to different situations, thereby confirming the **H2** hypothesis.

It can be concluded that the adaptation between the environments is not recipro-cal and environment and situations selection should be considered during training,
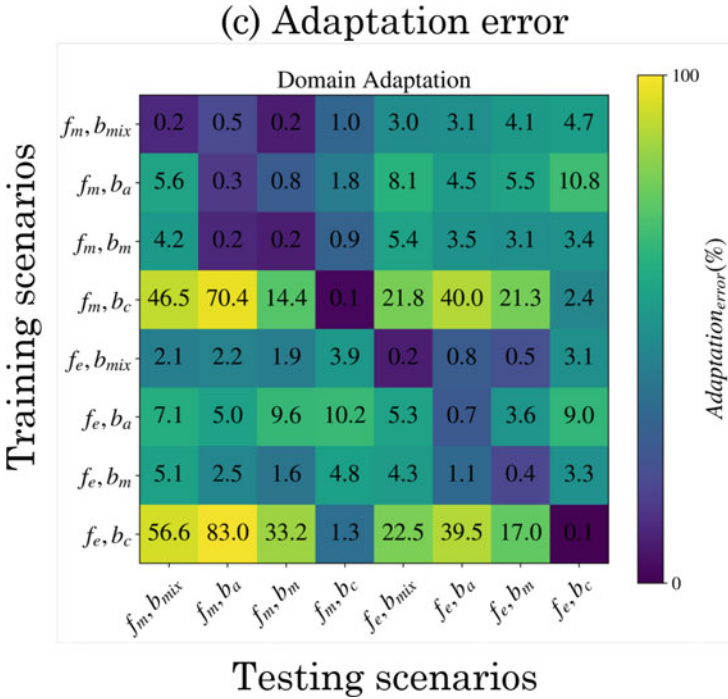
(c) Adaptation error

Fig. 13 The domain adaptation matrix with adaptation error (A_{error}) between different traffic scenarios and behaviors. The lower A_{error} the most suitable the adaptability between those domains

based on the application needs and target situations. The Domain adaptation matrix identifies the settings in which altruistic AVs can best learn cooperative policies that are robust to different traffic scenarios and human behaviors.

6.4.3   Transfer Learning

Through domain adaptation and transfer learning, we promote generalization while learning harder tasks efficiently from trained models and accelerate the learning process. We study how the policies learned during merging can be transferred to the exiting environment. For that, we train AVs agents from scratch for the mission/task of merging $AV_{merging}$ (T1), train AVs agents to drive on a highway, and then use that model as the starting point to learn the merging task $AV_{drive-to-merging}$ (T2), train AVs agents for the exiting task and then use that model as the starting point to learn the merging task $AV_{exiting-to-merging}$ (T3); and apply the same procedure for the exiting task, learning to exit from scratch $AV_{exiting}$ (T4), after learned how to drive $AV_{drive-to-exiting}$ (T5) and after learned how to merge $AV_{merging-to-exiting}$ (T6). The results of the experiments are presented in Fig. 14 and show that our transfer learning
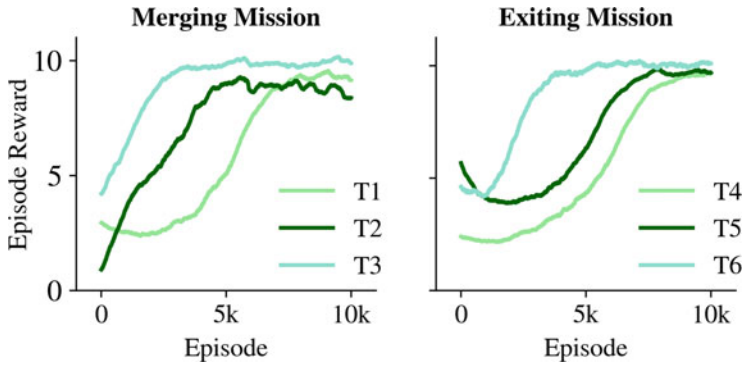
Fig. 14 The figure demonstrates how policies learned from merging can be transferred to the exiting environment to speed up the learning process while archiving similar performance to learning the task from scratch

approach speeds up the learning process while archiving similar performance as when learning the task from scratch.

6.4.4   Safety

Finally, we compared state of the art architectures related to our approach [10, 17, 23, 60] in terms of safety and efficiency to validate **H3**. We trained the different architectures in the same situations and examined their performance under different levels of HVs behaviors. As noted in Table 5 our safe altruistic agents consistently outperformed the other approaches (in bold is highlighted the best performance for each column), and the results are more notable when the level of aggressiveness is higher. We conclude that when using the safety prioritizer, immediate collisions are avoided reducing the overall number of crashes in the episodes. Our agents can learn from scratch not only how to drive, but also to understand the behavior of HVs and coordinate with them.

6.4.5   Importance of Social Coordination

We demonstrate that social awareness and coordination are essential to improve safety and reliability on the roads. Particularly in our sensitivity analyses (Fig. 9) we have shown that altruistic agents have a significant performance gain when compared to egoistic agents and the gain is more notable as the road becomes more aggressive. Additionally, to show that the performance gain vs. driver behaviors is not just because of a single altruistic agent but as a consequence of coordination among agents, we complement our results and conducted an experiment with the difference that only AV1 is altruistic and the others are egoistic AVs, we label this

Table 5 A comparison of the performance of related architectures. Our safe altruistic AVs outperform the other solutions, and performance improvements become more noticeable as the level of aggressiveness increases

Approaches	Aggressive HVs			Moderate HVs			Conservative HVs		
	C (%)	MF (%)	DT (m)	C (%)	MF (%)	DT (m)	C (%)	MF (%)	DT (m)
Conv2D+DQN [60]	31.2	28.9	316	25.4	20.3	302	14.0	7.9	274
Toghi et al. [17]	21.3	16.4	339	12.7	10.1	333	1.6	0.6	269
Conv3D+A2C [23]	14.8	12.6	341	9.4	8.8	328	1.1	0.1	267
Conv3D+DQN [10]	3.1	2.8	359	2.6	2.4	341	0.3	**0**	**284**
Ours	**0.2**	**0.1**	**397**	**0.1**	**0.1**	**354**	**0**	**0**	281

C Crashed, *MF* Mission Failed, *DT* Distance Traveled

Table 6 Importance of Social Coordination: AVs require to coordinate to enable a safe and seamless merging/exiting and none of them can achieve this goal if the others do not cooperate

	Aggressive HVs $C(\%)$	Moderate HVs $C(\%)$	Conservative HVs $C(\%)$
Multi-agent altruistic (MAA)	0.2%	0.1%	0%
Single altruistic agent (SAA)	24.1%	17.4%	2.3%

scenario as single altruistic agent **SAA**. Table 6 demonstrates the necessity of multi-agent coordination and the fact that a single altruistic AV, i.e., the Guide AV, is not able to achieve the mission of safe and seamless merging without help from the other AVs. Our results show that a non-cooperative SAA is not enough to guide the traffic and successful completion of the missions, as coordination is not guaranteed in a single-agent setting. All the AVs have to coordinate collectively to allow safe and efficient traffic, and this is unfeasible if the others do not collaborate. Table 6 complements our results in Fig. 9 and support the hypothesis **H1**.

6.5 Qualitative Analyses

We show a qualitative analysis of our altruistic AVs in the exit and merging scenarios. Figure 15 provides further intuition about the policies learned by altruistic AVs (*green*) in different situations, Figs. 15 and 16 show a set of snapshots for different policies learned in an exit/merging environment in the presence of HVs (*blue*) with different behaviors. In the presence of aggressive HVs, the guide AV has to slow down and guide the HVs in the platoon to allow a safe merging/exit of the mission vehicle; in this case, by slowing down the AV learn to compromise on their own utility for a more desirable social outcome. In the presence of moderate HVs
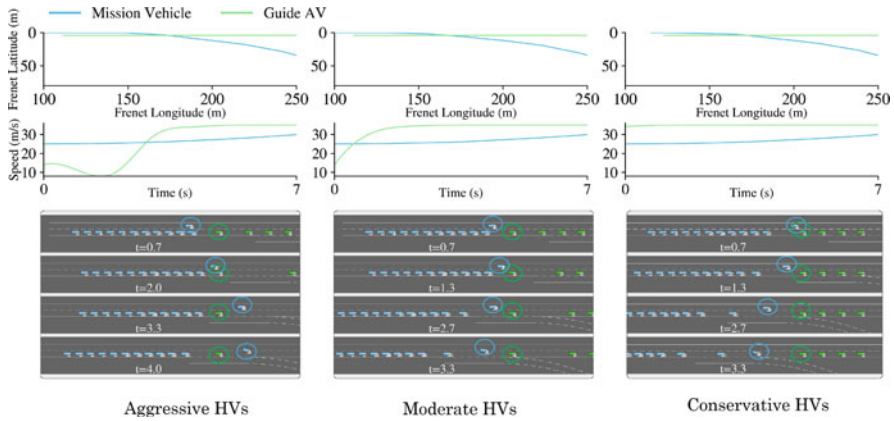
Fig. 15 Mission vehicle exiting the road under different HV behaviors (from left to right: aggressive, moderate and conservative HVs). AVs are shown in *green* and HVs are shown in *blue*
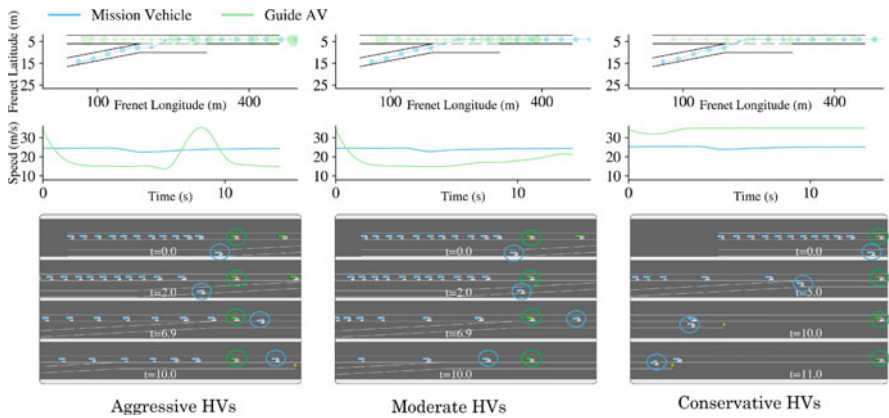

Fig. 16 Mission vehicle merging into the highway under different HV behaviors (from left to right: aggressive, moderate, and conservative HVs). AVs are shown in *green* and HVs are shown in *blue*. The diameter of the circles on the trajectory plot (first-row plot) shows the vehicles' speed

behaviors, the guide AV slows down (slowing down the vehicles in the platooning) to open a safe space for the mission vehicle and then quickly accelerates, the space created by the quick AV intervention is safe enough to allow the mission vehicle to exit/merge the road; in this case, the AV compromise in their own utility but does not need to compromise as much as in the aggressive traffic scenario, it learns to take sequences of actions to not only enable the mission vehicle to merge (by quick decelerating), but also manages to make the minimum compromise on its individual utility. Finally, in the conservative environment, the HVs are cautious enough to allow the mission vehicle to exit/merge safely, so the AVs learn to accelerate in those scenarios as the mission vehicle has enough space to merge, optimizing for

their own utility (higher speed and longer distance travel), while also considering other vehicles utilities and safety; in this case the AV doe not need to compromise their own utility, it learns that HV will allow the exiting/merging so AVs does not need to guide the traffic. It is important to notice that the policies are learned by AVs from experience to optimize the social utility, AVs learn to adapt to different scenarios and behaviors. Is interesting to observe that our AVs develop some form of social awareness and learn the HVs' behaviors from experience, acting accordingly to optimize traffic efficiency while prioritizing safety.

7 Concluding Remarks

AVs need to learn to co-exist with HVs vehicles as deploying egoistic AVs that solely account for their individual interests on the road leads to sub-optimal and non-desirable social outcomes. Social awareness and coordination are essential to improve safety and reliability on the roads. We demonstrate how altruistic AVs learn the decision-making process from experience, considering the interests of all vehicles while prioritizing safety and optimizing a general decentralized social utility function. We expose the settings for our MARL problem in which transfer learning and domain adaptation are more feasible, and conducted a sensitivity analysis under different HVs' behaviors. Our experiments reveal that altruistic AVs learn to leverage social coordination to improve safety and reliability. Our social-aware AVs are robust to heterogeneous driver behaviors and can form alliances and affect the behavior of HVs to create socially-desirable outcomes that benefit the group of the vehicles.

Future Work Although we explored various elements of social navigation in a variety of settings and the presence of diverse HV behaviors, the HV models used are not from real human driver data, and the traffic scenarios are limited to merging and exiting. However, we believe that by leveraging and learning from actual human data and traffic circumstances, our approach might be beneficial in practical traffic conditions. For this strategy to be used in real-world circumstances, more attention to safety is necessary. We intend to investigate more sophisticated architectures and state representations in future work, as well as develop a more realistic simulation environment that incorporates data from real-world traffic and can handle more complex interactions between HVs and AVs, as well as diverse traffic agents like bicycles and pedestrians. Despite the drawbacks, we are excited to see safe and reliable social-aware AVs on the road that learns from experience. Beyond driving, we expect these principles to be applied to general multi-agent human-robot interactions in which agents influence humans and collaborate safely for a socially beneficial result.

References

1. Cosgun, A., Ma, L., Chiu, J., Huang, J., Demir, M., Anon, A.M., Lian, T., Tafish, H., Al-Stouhi, S.: Towards full automated drive in urban environments: a demonstration in gomentum station, California. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 1811–1818. IEEE (2017)
2. Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., Rus, D.: Social behavior for autonomous vehicles. Proc. Natl. Acad. Sci. **116**(50), 24972–24978 (2019)
3. Sagberg, F., Selpi, Piccinini, G.F.B., Engström, J.: A review of research on driving styles and road safety. Human Factors **57**(7), 1248–1275 (2015)
4. Toghi, B., Saifuddin, M., Mughal, M., Fallah, Y.P.: Spatio-temporal dynamics of cellular v2x communication in dense vehicular networks. In: 2019 IEEE 2nd Connected and Automated Vehicles Symposium (CAVS), pp. 1–5. IEEE (2019)
5. Shah, G., Valiente, R., Gupta, N., Gani S.O., Toghi, B., Fallah, Y.P., Gupta, S.D.: Real-time hardware-in-the-loop emulation framework for dsrc-based connected vehicle applications. In: 2019 IEEE 2nd Connected and Automated Vehicles Symposium (CAVS), pp. 1–6. IEEE (2019)
6. Valiente, R., Zaman, M., Ozer, S., Fallah, Y.P.: Controlling steering angle for cooperative self-driving vehicles utilizing cnn and lstm-based deep networks. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2423–2428. IEEE (2019)
7. Razzaghpour, M., Shahram, S., Valiente, R., Fallah, Y.P.: Impact of communication loss on mpc based cooperative adaptive cruise control and platooning. Preprint (2021). arXiv:2106.09094
8. Valiente, R., Raftari, A., Zaman, M., Fallah Y.P., Mahmud, S.: Dynamic object map based architecture for robust cvs systems. SAE Technical Paper, Technical Report, 2020
9. Aoki, S., Higuchi, T., Altintas, O.: Cooperative perception with deep reinforcement learning for connected vehicles. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 328–334. IEEE (2020)
10. Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., Fallah, Y.P.: Cooperative autonomous vehicles that sympathize with human drivers. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2021)
11. Jami, A., Razzaghpour, M., Alnuweiri, H., Fallah, Y.: Augmented driver behavior models for high-Fidelity simulation study of crash detection algorithms. https://arxiv.org/pdf/2208.05540.pdf
12. Harris P.B., Houston J.M., Vazquez J.A., Smither J.A., Harms, A., Dahlke, J.A., Sachau, D.A.: The prosocial and aggressive driving inventory (padi): a self-report measure of safe and unsafe driving behaviors. Accid. Anal. Prev. **72**, 1–8 (2014)
13. Vallières, E.F., Vallerand, R.J., Bergeron, J., McDuff, P.: Intentionality, anger, coping, and ego defensiveness in reactive aggressive driving. J. Appl. Soc. Psychol. **44**(5), 354–363 (2014)
14. Bouton, M., Nakhaei, A., Fujimura, K., Kochenderfer, M.J.: Cooperation-aware reinforcement learning for merging in dense traffic. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 3441–3447. IEEE (2019)
15. Sadigh, D., Sastry, S., Seshia S.A., Dragan, A.D.: Planning for autonomous cars that leverage effects on human actions. In: Robotics: Science and Systems, vol. 2. Ann Arbor (2016)
16. Wu, C., Kreidieh, A., Vinitsky, E., Bayen, A.M.: Emergent behaviors in mixed-autonomy traffic. In: Conference on Robot Learning, pp. 398–407. PMLR (2017)
17. Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., Fallah, Y.P.: Social coordination and altruism in autonomous driving. Preprint (2021). arXiv:2107.00200
18. Rios-Torres, J., Malikopoulos, A.A.: A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps. IEEE Trans. Intell. Trans. Syst. **18**(5), 1066–1077 (2016)
19. Mahjoub, H.N., Raftari, A., Valiente, R., Fallah Y.P., Mahmud, S.K.: Representing realistic human driver behaviors using a finite size gaussian process kernel bank. In: 2019 IEEE Vehicular Networking Conference (VNC), pp. 1–8. IEEE (2019)
20. Li, Z., Kalabić, U., Chu, T.: Safe reinforcement learning: Learning with supervision using a constraint-admissible set. In: 2018 Annual American Control Conference (ACC), pp. 6390–6395. IEEE (2018)

21. Lin, Y., McPhee, J., Azad, N.L.: Anti-jerk on-ramp merging using deep reinforcement learning. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 7–14. IEEE (2020)
22. Sadigh, D., Landolfi, N., Sastry S.S., Seshia S.A., Dragan A.D.: Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state. Auton. Robot. **42**(7), 1405–1426 (2018)
23. Toghi, B., Valiente, R., Sadigh, D., Pedarsani, R., Fallah, Y.P.: Altruistic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2021)
24. Foerster J.N., Chen R.Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., Mordatch, I.: Learning with opponent-learning awareness. Preprint (2017). arXiv:1709.04326
25. Xie, A., Losey, D., Tolsma, R., Finn, C., Sadigh, D.: Learning latent representations to influence multi-agent interaction. In: Proceedings of the 4th Conference on Robot Learning (CoRL) (2020)
26. Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., Whiteson, S.: Counterfactual multi-agent policy gradients. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
27. Egorov, M.: Multi-agent deep reinforcement learning. CS231n: Convolutional Neural Networks for Visual Recognition, pp. 1–8 (2016)
28. Omidshafiei, S., Pazis, J., Amato, C., How J.P., Vian, J.: Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In: International Conference on Machine Learning, pp. 2681–2690. PMLR (2017)
29. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. Preprint (2017). arXiv:1706.02275
30. Brown, K., Driggs-Campbell, K., Kochenderfer, M.J.: A taxonomy and review of algorithms for modeling and predicting human driver behavior. arxiv e-prints, article. Preprint (2020). arXiv:2006.08832
31. Ivanovic, B., Schmerling, E., Leung, K., Pavone, M.: Generative modeling of multimodal multi-human behavior. In: RSJ International Conference on Intelligent Robots and Systems, pp. 3088–3095. IEEE (2018)
32. Lauer, M., Riedmiller, M.: An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In: In Proceedings of the Seventeenth International Conference on Machine Learning. Citeseer (2000)
33. Toghi, B., Grover, D., Razzaghpour, M., Jain, R., Valiente, R., Zaman, M., Shah, G., Fallah, Y.P.: A maneuver-based urban driving dataset and model for cooperative vehicle applications. In: 2020 IEEE 3rd Connected and Automated Vehicles Symposium (CAVS), pp. 1–6. IEEE (2020). https://ieeexplore.ieee.org/document/9334665
34. Chandra, R., Bhattacharya, U., Mittal, T., Bera, A., Manocha, D.: Cmetric: A driving behavior measure using centrality functions. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2035–2042. IEEE (2020)
35. Constantinescu, Z., Marinoiu, C., Vladoiu, M.: Driving style analysis using data mining techniques. Int. J. Comput. Commun. Control **5**(5), 654–663 (2010)
36. Beck K.H., Ali, B., Daughters, S.B.: Distress tolerance as a predictor of risky and aggressive driving. Traffic Inj. Prev. **15**(4), 349–354 (20140
37. Pokle, A., Martín-Martín, R., Goebel, P., Chow, V., Ewald, H.M., Yang, J., Wang, Z., Sadeghian, A., Sadigh, D., Savarese, S., et al.: Deep local trajectory replanning and control for robot navigation. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 5815–5822. IEEE (2019)
38. Kuderer, M., Gulati, S., Burgard, W.: Learning driving styles for autonomous vehicles from demonstration. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 2641–2646. IEEE (2015)
39. Hadfield-Menell, D., Russell S.J., Abbeel, P., Dragan, A.: Cooperative inverse reinforcement learning. Adv. Neural Inf. Proces. Syst. **29**, 3909–3917 (2016)

40. Trautman, P., Krause, A.: Unfreezing the robot: navigation in dense, interacting crowds. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 797–803. IEEE (2010)
41. Nikolaidis, S., Ramakrishnan, R., Gu, K., Shah, J.: Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 189–196. IEEE (2015)
42. Wu, C., Bayen A.M., Mehta, A.: Stabilizing traffic with autonomous vehicles. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 6012–6018. IEEE (2018)
43. Lazar, D.A., Bıyık, E., Sadigh, D., Pedarsani, R.: Learning how to dynamically route autonomous vehicles on shared roads. Preprint (2019). arXiv:1909.03664
44. Bıyık, E., Lazar, D.A., Pedarsani, R., Sadigh, D.: Incentivizing efficient equilibria in traffic networks with mixed autonomy. IEEE Trans. Control Netw. Syst. **8**(4), 1717–1729 (2021)
45. Li, S., Yan, Z., Wu, C.: Learning to delegate for large-scale vehicle routing. Adv. Neural Inf. Process. Syst. **34** (2021)
46. Hickert, C., Li, S., Wu, C.: Cooperation for scalable supervision of autonomy in mixed traffic, pp. arXiv–2112. e-prints (2021)
47. Razzaghpour, M., Mosharafian, S., Raftari, A., Mohammadpour Velni, J., and Fallah, Y.P.: Impact of information flow topology on safety of tightly-coupled connected and automated vehicle platoons utilizing stochastic control. In: ECC (2022)
48. Wang W.Z., Beliaev, M., Biyik, E., Lazar D.A., Pedarsani, R., Sadigh, D.: Emergent prosociality in multi-agent games through gifting. In 30th International Joint Conference on Artificial Intelligence (IJCAI) (2021)
49. Wang, J., Zhang, Q., Zhao, D., Chen, Y.: Lane change decision-making through deep reinforcement learning with rule-based constraints. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2019)
50. Nageshrao, S., Tseng H.E., Filev, D.: Autonomous highway driving using deep reinforcement learning. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 2326–2331. IEEE (2019)
51. Mohammadhasani, A., Mehrivash, H., Lynch, A., Shu, Z.: Reinforcement learning based safe decision making for highway autonomous driving. Preprint (2021). arXiv:2105.06517
52. Chen, D., Li, Z., Wang, Y., Jiang, L., Wang, Y.: Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic. Preprint (2021). arXiv:2105.05701
53. Le, V.-A., Malikopoulos, A.A.: A cooperative optimal control framework for connected and automated vehicles in mixed traffic using social value orientation. Preprint (2022). arXiv:2203.17106
54. Murphy, R.O., Ackermann, K.A.: Social preferences, positive expectations, and trust based cooperation. J. Math. Psychol. **67**, 45–50 (2015).
55. Garapin, A., Muller, L., Rahali, B.: Does trust mean giving and not risking? Experimental evidence from the trust game. Rev. Econ. Polit. **125**(5), 701–716 (2015)
56. Müller, L., Risto, M., Emmenegger, C.: The social behavior of autonomous vehicles. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, Ser. UbiComp '16, pp. 686–689. Association for Computing Machinery, New York (2016) [Online]. Available: https://doi.org/10.1145/2968219.2968561
57. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E **62**(2), 1805 (2000)
58. Kesting, A., Treiber, M., Helbing, D.: General lane-changing model mobil for car-following models. Transp. Res. Rec. **1999**(1), 86–94 (2007)
59. Leurent, E., Blanco, Y., Efimov, D., Maillard, O.-A.: Approximate robust control of uncertain dynamical systems. Preprint (2019). arXiv:1903.00220
60. Van Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1 (2016)