Robustness and Adaptability of Reinforcement Learning based Cooperative Autonomous Driving in Mixed-autonomy Traffic

Rodolfo Valiente[1], Behrad Toghi[1], Ramtin Pedarsani[2], Yaser P. Fallah[1]

Building autonomous vehicles (AVs) is a complex problem, but enabling them to operate in the real-world where they will be surrounded by human-driven vehicles (HVs) is extremely challenging. Prior works have shown the possibilities of creating inter-agent cooperation between a group of AVs that follow a social utility. Such altruistic AVs can form alliances and affect the behavior of HVs to achieve socially-desirable outcomes. We identify two major challenges in the co-existence of AVs and HVs. First, social preferences and individual traits of a given human driver, e.g., selflessness, and aggressiveness are unknown to an AV, and it is almost impossible to infer them in real-time during a short AV-HV interaction. Second, contrary to AVs that are expected to follow a policy, HVs do not necessarily follow a stationary policy and therefore are extremely hard to predict. To alleviate the above-mentioned challenges, we propose a multi-agent reinforcement learning solution for altruistic AVs that robustifies them to different human behaviors and also constrains AVs to a safe action space. We investigate the robustness and sensitivity of AVs to various HVs behavioral traits and present the settings in which our AVs can learn cooperative policies that are adaptable to different situations.

Index Terms—**Behavior Planning, Cooperative Driving, Mixed-autonomy, Reinforcement Learning, Robustness**

I. INTRODUCTION

THE development of autonomous vehicles (AVs) is on the verge of passing beyond the laboratory and simulation tests and is shifting towards addressing the challenges that limit their practicality to be used in today's society. While there is still need for further technological improvements to enable safe and smooth operation of a single AV, a great deal of research attention is being focused on the emerging challenge of operating multiple AVs and the co-existence of AVs and human-driven vehicles (HVs) [1], [2]. A realistic outlook for the adoption of autonomous vehicles on our roads is a mixed-traffic scenario in which human drivers with different driving styles and social preferences share the road with AVs that are perhaps built by different manufacturers and hence follow different policies [3], [4]. In this work, we seek a solution that can ensure the safety and robustness of AVs in the presence of human drivers with heterogeneous behaviorial traits.

We start with identifying the major challenges in the domain of behavior planning and prediction for AVs in mixed-autonomy traffic. As a preliminary, it is important to distinguish between the individual traits of a human driver, e.g., aggressiveness,conservativeness, risk-tolerance, and their social preferences, e.g., egoism and altruism. Despite the correlation between the two categories, they arise from different natures and also lead to different behaviors in mixed traffic. As an example, an aggressive driver is not necessarily egoistic and selfish, but their aggression might limit their capability to cooperate with other drivers and taking part in a socially-desirable co-existence with AVs [5]–[7]. As mentioned before, human drivers are heterogeneous in their individual traits and social preferences, which makes the autonomous vehicle behavior planning extremely difficult, as it is challenging for the AV to predict the type of behavior it is going to face when dealing with a human driver. Additionally, relying on real-time inference of HVs' behaviors is not always viable as the interaction time between vehicles can be short-lived, e.g., two vehicles that meet in an intersection.

In a pursuit to alleviate the challenges of this co-existence and enabling social navigation for AVs, existing works either rely on models of human behavior derived from pre-recorded driving datasets [8], [9] or defining social utilities that can enforce a cooperative behavior among AVs and HVs [2], [10]. The majority of the existing literature relies on simulated environments or human-in-the-loop simulations, which limits the capabilities of modeling the interactions of human drivers with AVs and implementing the heterogeneity of human behaviors. This shortcoming hinders the applicability of the resulting solutions as they are often limited to the human behaviors with which they have interacted during the training. In order to accommodate for this, some of the proposed policies in the literature take an overly-conservative approach when interacting with humans [11]. Not only this approach leaves the AVs vulnerable to other aggressive drivers, especially in competitive scenarios such as intersections, but also can cause traffic congestion and potential safety threats [1], [12].

We build on our prior work in [3] and aim to develop a safe and robust training regimen that enables the AVs to work together and influence the behavior of human drivers to create socially-desirable traffic outcomes, regardless of humans' driving individual traits and social preferences. Furthermore, we emphasize on the importance of safety in social settings and constrain the AVs' policies to remove high-risk actions that can cause safety threats. Our work in this paper is built on the following key insights. First, we rely on learning from experience in a decentralized reinforcement learning framework that optimizes for a social utility, and expose the learning agents to a wide spectrum of driver behaviors. By doing so, the resulting agents become robust to the behavior of human drivers and are able to handle cooperative-competitive behaviors regardless of HV's level of aggression and social preference. Second, a safety prioritizer is proposed to avoid high-risk actions that can

*This material is based upon work partially supported by the National Science Foundation under Grant No. CNS-1932037.

[1] Rodolfo Valiente, Behrad Toghi, and Yaser P. Fallah are with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, USA. rvalienter90@knights.ucf.edu

[2] Ramtin Pedarsani is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA.

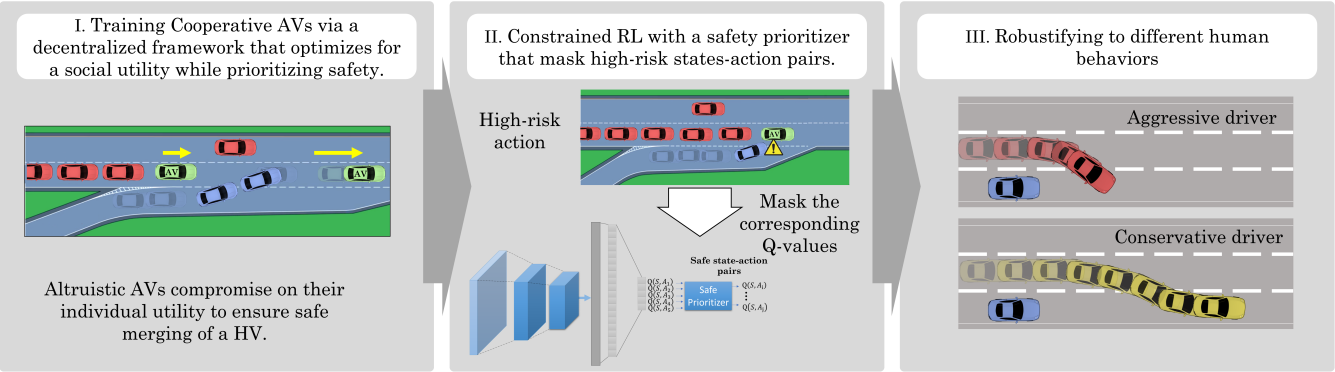

Fig. 1: An overview of our methodology for ensuring the safety and robustness of cooperative autonomous vehicles in interaction with human-driven vehicles.

undermine driving safety. The overview of our methodology is presented in Figure 1.

Ultimately, the focus of this paper is on exploring the safe and robust decision-making problem in a mixed-autonomy MARL environment, in inherently competitive driving scenarios, such as the ones illustrated in Figure 2, where cooperation is required for safety and efficiency. The purpose is not to fully solve the autonomous driving problem, but instead, use it as an example to investigate the effectiveness of societal concepts from psychology literature within the MARL domain. Further work is required to use these ideas on real roads. Nevertheless, we are encouraged to see altruistic AVs that are safe, robust, and can learn to influence humans in a desired way without the limitations of current solutions.

Our main contributions are:

- We begin by formulating the mixed-autonomy traffic as a stochastic game and introduce a general decentralized framework for training cooperative AV, that optimizes for a social utility while prioritizing safety and allowing adaptability.
- A novel training regimen is introduced that robustifies the AVs' capability in creating socially-desirable outcomes with regards to human drivers' behavior.
- We proposed a safety prioritizer that constrained the policy of cooperative AVs to ensure the safety of their behavior via masking the Q-states that lead to high-risk outcomes.

II. Literature Review

MARL has received a lot of attention from the research community in recent years. MARL algorithms that assume separately trained agents perform poorly due to the intrinsic non-stationarity of the environment [13]. Some efforts to solve this issue assume that all agents observe the global state [14] or that they can share their states with their neighbors [15]. These assumptions address the non-stationarity challenge; however, they are unfeasible on real roads [16]. To address this challenge, [17] propose a centralized critic that reduce the influence of non-stationarity during the learning process. Xie *et al.* consider a MARL approach that learns latent representations of the agent's policies, modeling agent strategies that depend on long interactions, alleviating the non-stationary effect, and enabling better performance and co-adaptation [18]. In [19] is further investigated the impact of interactions on agent's modeling. Authors in [20] present a RL agent that learns to acquire social norms from public sanctions using a decentralized framework.

A. Driver Behavior and Social Navigation

Social navigation in mixed autonomy has shown the potential of collaboration among AVs and HVs [21]. Current works in social navigation tackle the MARL cooperation by assuming the nature of agent interactions [22], [23] or by directly modeling or classifying human driver behaviors [24], [25]. Different methods to predict or classify driver behaviors are based on driver attributes [26], graph theory [27], game theory [1] and data mining [28]. Authors in [25] present an approach to modeling and predicting human behavior in situations with several humans interacting in highly multi-modal scenarios that could allow AVs to predict human reactions. [24] can be referred for a comprehensive study on modeling and prediction in multi-agent traffic scenarios.

In [29] and [4] the driving patterns of humans are learned from demonstration through inverse RL. An approach based on imitation learning is proposed by [8] to learn a reward function for human drivers and demonstrates how AVs can manipulate human behaviors. In [30] a centralized game-theory model for cooperative inverse reinforcement learning is proposed. Several works take a more abstract and traffic-level perspective [31]–[33]. Differently, we rely on implicitly learning from experience altruistic behaviors that facilitate AVs coordination without the need for a human model or counting on their collaboration.

B. Safe and Robust Decision-Making

In addition to the socially advantageous behavior of altruistic AVs, it is important to consider robustness and safety. Safety is essential for AVs and is particularly important for AVs trained using RL. In that direction, several works take a pure reward shaping approach to avoid collisions. While this is a common practice in RL, safety is not implicitly prioritized

and AVs using those RL algorithms may not behave safely in some scenarios, as the agents could choose dangerous actions due to function approximation. To address this challenge, the idea of safe RL is proposed in [11] improving safety in unseen driving environments in which the RL algorithm behaves unsafely. In [34] is presented a rule-based decision-making framework that examines the trajectories given by the controller and replaces the actions causing collisions. Nageshrao *et al.* [35] includes a short-horizon safety supervisor to substitute risky actions with safer ones. Nevertheless, these studies consider oversimplified and non-realistic environments. The work in [36], [37] utilizes a Q-masking approach to prevent collisions, removing the actions that could result in a crash. Chen *et al.* present a novel priority-based safety supervisor to significantly reduce collisions [38]. In this work, we leverage this idea to improve the safety of the altruistic agents while also training the cooperative agents to be robust to different driver behaviors and scenarios using a decentralized reward function, local actions, and assuming partial observability.

III. Problem Formulation

We can formalize the MARL problem as a centralized or decentralized problem. A centralized controller that assigns a central joint reward and joint action is straightforward. Nevertheless, such assumptions are impractical in the real-world. Therefore, we focus on a decentralized controller where agents have partial observability and consequently, we formulated the problem as a partially observable stochastic game (POSG) defined by $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\{1,\dots,N\}}, \{\mathcal{O}_i\}_{i\in\{1,\dots,N\}}, P, \{R_i\}, \gamma\rangle$ 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).
- $\mathcal{A}_i$: a set of possible actions for agent i.
- $\mathcal{O}_i$: a set of observations for agent i.
- P: a state transition probability function from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$, $P(S = s'|S = s, A = a)$.

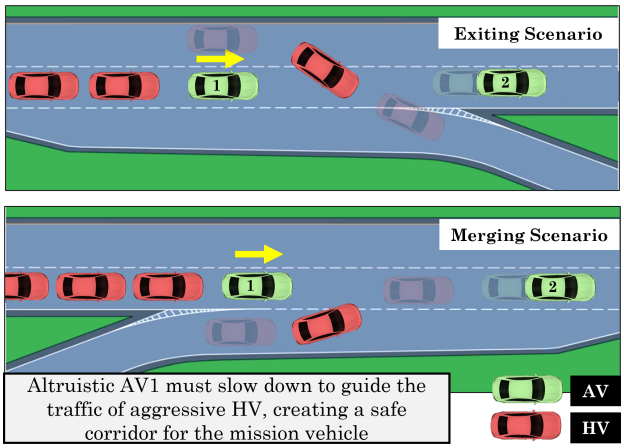

Fig. 2: Road shared by AVs (green) and aggressive HVs (red). Altruistic AVs must learn to coordinate to allow for a safe and efficient merging/exiting while also being robust to different scenarios and behaviors and ensuring safety in decision-making.

- R_i: a reward function for the i^{th} agent, $R_i(s_i, a_i)$.
- γ: a discount factor, $\gamma \in [0, 1]$.

Our agents have no access to the exact environmental state but only a local observation which is correlated with the state, increasing the difficulty of solving the POSG. The POSG can be described as follows: at every time step t, s_t is the state of the environment, each agent senses the environment and obtains a local observation $o_i : \mathcal{S} \to \mathcal{O}_i$, based on o_i and its stochastic policy π_i, it selects an action from the action space $a_i \in \mathcal{A}_i$. As a result, the agent moves to the next state s_i' and obtains a decentralized reward $r_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$. 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 reward over an infinite time horizon, and find the corresponding optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}$.

An optimal policy maximizes the action-value function, i.e., $\pi^*(s) = \arg\max_a Q^*(s, a)$ where $Q^\pi(s, a) := \mathbb{E}[\sum_{i=1}^{\infty} \gamma^i R(s_i, \pi(s_i))|s_0 = s, a_0 = a]$. The optimal action-value function is determined by solving the Bellman equation

$$Q^*(s, a) = \mathbb{E}_{s'\sim P(.|s,a)}\left[R(s, a) + \gamma \max_{a'} Q^*(s', a')\right] \quad (1)$$

A. Double Deep Q-Network (DDQN)

We use Double Deep Q-Network (DDQN) as the function approximator to estimate the action-value function, i.e., $\tilde{Q}(.; \mathbf{w}) \cong Q(.)$ (with weights $\mathbf{w}$) [39]. DDQN improves Deep Q-Network (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

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

$$Target^{DDQN} = R(s, a) + \gamma\tilde{Q}(s', \arg\max_{a'}\tilde{Q}(s', a'; \mathbf{w}); \hat{\mathbf{w}})) \quad (3)$$

The DDQN algorithm then applies mini-batch gradient descent as $\mathbf{w}_{i+1} = \mathbf{w}_i + \alpha_i\hat{\nabla}_{\mathbf{w}}\mathcal{L}(\mathbf{w}_i)$, 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 $\mathbf{w}_i$ and $\hat{\mathbf{w}}$ are the weights of the target network which are updated at a lower frequency ($Target_{update}$).

B. Driving Scenarios

Our goal is to study driving scenarios where the absence of coordination by the AVs compromises safety and efficiency. Additionally, we aim to investigate adaptability among competitive scenarios and driver behaviors. For this purpose, we define a set of scenarios $\mathcal{F}$ and choose highway merging and exiting ramp as the base scenarios where a mission vehicle (merging/exiting) attempts to complete its task in a mixed-traffic environment as in Figure 2. The merging and exiting scenario are defined as $f_m, f_e \in \mathcal{F}$ respectively. We select such situations because of their intrinsic closeness and competitive characteristics, since the merging/exiting vehicle's local utility conflicts with that of the highway vehicles.

C. Social Value Orientation and Altruistic AVs

Social Value Orientation (SVO) characterizes the individual's preference to account for the interests of others vs. their own interest [1]. The behavior of a human or an AV can fluctuate from egoistic to absolutely altruistic based on the importance given to the utility of others. The SVO of humans is uncertain, therefore we depend on AVs instead to guide the traffic toward more socially advantageous goals. Formally, the SVO angle ϕ of an AV, determines how the AV balances its own benefit against that of others. In terms of rewards, we can define the total reward R_i of an AV as: $R_i(s_i, a_i) = \cos\phi_i r_i^{ego} + \sin\phi_i r_i^{social}$, in which r_i^{ego} is the AV's specific reward (egoistic) and r_i^{social} is the overall reward of other vehicles (social) respect to the i^{th} AV [2], [3]. The SVO angle can be changed from $\phi = 0$ (purely egoistic) to $\phi = \pi/2$ (purely altruistic). Nevertheless, none of the two extremes is optimum, and a point in between yields the most socially advantageous result, defined as the optimal SVO angle ϕ^*. SVO helps explain the behaviors that allow the mission vehicle to merge or exit in Figure 2. Without SVO, the mission vehicle in Figure 2 could cause traffic congestion or an unsafe situation. AVs need to consider SVO, since HVs can not communicate that directly, and we should not expect HVs to cooperate.

D. Driving Behaviors

The challenge of simulating diverse behaviors can be framed as the problem of obtaining the suitable range of parameters that can generate the heterogeneous behaviors within the simulator. Many studies from social traffic psychology establish that driving behavior falls between aggressive and conservative. Nonetheless, the precise definitions differ between studies [5]. In general, the term "aggressive driving" covers a range of unsafe driving behaviors like overspeeding or running red lights. However, the causes of aggressive driving come in various forms and are not always obvious. Some are undesirable roadway situations, while others are individual traits or states of mind. Furthermore, there is a relationship between aggressiveness and egoism, as egoistic drivers usually do not yield and also tend to engage in speeding, risk-taking, and similar aggressive behaviors. While there is a correlation between these terms [5]–[7], for the purpose of this paper, we separate egoism from aggressiveness by characterizing social preferences and individual traits.

We differentiate social preferences and individual traits of HVs as they lead to different behaviors. First, we characterize egoism and altruism as social preferences, and identify an egoistic HV as a selfish driver who accounts for its own utility independently of its aggressiveness. Second, we characterize aggressiveness and conservativeness as individual traits, and identify an aggressive HV as a driver whom the outcome of their actions causes aggressive behaviors. Social preferences as egoism are characterized by their social goals and intentions, whereas individual traits as aggressiveness are characterized by the consequences of their actions. For instance, a driver could be egoistic and conservative. We could imagine a driver who drives cautiously in order to protect himself (selfish motivation/preference) and, as a consequence, behaves conservatively (outcome of his actions).

Formally, in our simulation, social preferences (egoism or altruism) are characterized by the AV's SVO angular preference ϕ; and individual traits (aggressiveness, conservativeness, etc) by the HV driver model parameters ($\mathcal{P}$) as described in section IV-D. Based on the values of these parameters, a vehicle will exhibit aggressive or conservative behaviors. In our experiments, we assume the SVO of HVs to be unknown as they can not 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 section IV-D.

E. Problem Formulation

We formulate the problem as the POSG defined; where the road is shared by a set of HVs $h_i \in \mathcal{H}$, with an undetermined SVO ϕ_i and heterogeneous behaviors $b_i \in \mathcal{B}$; a set 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 merging/exiting) and can be 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 cooperative optimal 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 sensitivity 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 $\tilde{\mathbf{o}}_i \in \widetilde{\mathcal{O}}_i$, 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 AVs that learn how to drive in a mixed-autonomy scenario in a robust, efficient and safe manner while benefiting all the vehicles on the road.

IV. Safe and Robust Altruistic Driving

To drive in a mixed-autonomy environment in a robust and safe manner, we propose a MARL approach with a general decentralized reward function that optimizes for a social utility by inducing altruism in the agents; the general reward accounts for any anticipated vehicle's mission, allowing it to be applied to different scenarios and tasks; and ensuring safety by adding a safety prioritizer. We train altruistic AVs that learn from experience to perform a task, account for the interests of all the vehicles, while being able to adapt to other traffic situations safely. We carefully design an appropriate action and observation space, a decentralized general reward function, a suitable architecture, and a safety prioritizer to promote the desired safe altruistic behavior in AVs' decision-making process.

Action Space. We define a high-level action space $\mathcal{A}$ of discrete meta-actions for decision-making. In particular, we select a set of five high-level actions as $a_i \in \mathcal{A}_i = [$`Change to Right Lane`, `Change to Left Lane`, `Accelerate`, `Decelerate`, `Idle`$]^T$. 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* semantic state representation that embeds the relative speed of the vehicle with respect to the ego vehicle in pixel values, as in [2]. We represent the information in multiple semantic channels that embed: 1) the AVs, 2) the HVs , 3) the mission vehicle, 4) an attention map to highlight the position of the ego vehicle, and 5) the road layout. To map into pixels the relative speed of the vehicles, we use a clipped logarithmic function which improves the dynamic range and shows better results than a straightforward linear mapping. As temporal information is necessary for safe decision-making, we use a history of VelocityMaps successive observations.

A. Decentralized General Reward

We train the AVs from scratch using local observations and a decentralized reward structure and expect them to learn the driving task in different scenarios while accounting for individual diver's missions. Consequently, we design a well-engineered general reward function that accounts for the social utility, traffic metrics and desired missions. The agent's $I_i \in \mathcal{I}$ local reward is defined as

$$\begin{aligned} R_i(s_i, a_i) &= R^{\text{ego}} + R^{\text{social}} \\ R^{\text{ego}} &= \cos\phi_i r_i(s_i, a_i) \\ R^{\text{social}} &= \sin\phi_i \Big[\sum_j r^{\text{AV}}_{i,j}(s_i, a_i) + \sum_j r^{M}_{i,j}(s_i, a_i) \\ &\quad + \sum_k r^{\text{HV}}_{i,k}(s_i, a_i) + \sum_k r^{M}_{i,k}(s_i, a_i) \Big] \end{aligned} \tag{4}$$

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. We decouple the social component in cooperation (the altruistic behavior among AVs) and sympathy (AV's altruism toward HVs) as they differ in nature. The sympathy term, $r^{\text{HV}}_{i,k}$ considers the individual reward of the HVs, while the cooperation term, $r^{\text{AV}}_{i,j}$ the individual reward of the other AVs, and are defined as

$$r^{\text{HV}}_{i,k} = \frac{\mathcal{W}_k}{d^{\lambda}_{i,k}} \sum_m \omega_m x_m \quad r^{\text{AV}}_{i,j} = \frac{\mathcal{W}_j}{d^{\lambda}_{i,j}} \sum_m \omega_m x_m \tag{5}$$

in which $d_{i,k}/d_{i,j}$ represent the distance between the agent and the corresponding HV/AV, λ is a dimensionless coefficient, $\mathcal{W}_k$ a weight value for individual vehicle's importance, m are the traffic metrics 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.

$$r^{\text{M}}_{i,j} = \begin{cases} \frac{w_j}{(d_{i,j})^{\mu}}, & \text{if} f(j) \\ 0, & \text{o.w.} \end{cases} \quad r^{\text{M}}_{i,k} = \begin{cases} \frac{w_k}{(d_{i,k})^{\mu}}, & \text{if} f(k) \\ 0, & \text{o.w.} \end{cases} \tag{6}$$

The function $f(v)$ is an independent function to evaluate the mission; $f(v)$ return 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 individual vehicle's mission (importance of the mission). This allows to define a general reward independent of the driving scenario and mission goals for different vehicles. In our experiments, a **HV** can be assigned a merging mission or a highway exiting mission, refer to Figure 2.

B. Deep MARL architecture for Cooperative Driving

We use a 3D Convolutional Neural Network (CNN) with a safety prioritizer as presented in Figure 3. The 3D CNN acts as a feature extractor and uses a history of VelocityMap observation to account for the temporal information.

To tackle the non-stationarity of MARL, we train the agents in a semi-sequential approach, as in [2]. 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 sample efficiency and train the agent safely, reducing episode resets due to imminent collisions, we use a safety prioritizer that when the action selected by the agent policy is unsafe, it selects a safe action, and store the unsafe action (a_t) and the related observation in the RM with a suitable penalty on the reward (r_{unsafe}) for the unsafe state-action pair. Those pairs are not removed so the agent can also learn from unsafe experiences. The experience $(\tilde{\mathbf{o}}_t, a_t, r_{unsafe}, \emptyset)$ is stored in RM with a terminal next state $\emptyset$, the target for this unsafe pair (s, a_t) is $Target(s, a_t)^{DDQN} = r_{unsafe}$. The details of the safety prioritizer are given in the next section IV-C. **Algorithm** 1 summarizes the overall methodology of our safety prioritized deep MARL architecture. Additionally, we do not initiate the learning process until the replay buffer is filled with a minimum number of sample simulations.

C. Safety prioritizer

As safety is an essential requirement in autonomous navigation, we add a safety prioritizer to the MARL algorithm, to avoid and penalize imminent collisions. This allows the agent to improve sample efficiency during training and avoid collisions during deployment. If the agent encounters an unseen scenario and decides to take an unsafe action, that action will be avoided. The safety prioritizer improves the simulation results and is critical in real-life situations. The safety prioritizer consists of **Algorithm 2** and **Algorithm 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

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 state s_1 and local observation $\tilde{\mathbf{o}}_1 = f(s_1)$
  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 $a_t = \max_{a' \in A} Q(\tilde{\mathbf{o}}_t, a'; \mathbf{w})$
        if a_t is unsafe (**Algorithm 2**) **then**
          Store experience $(\tilde{\mathbf{o}}_t, a_t, r_{unsafe}, \emptyset)$ in RM
          a_t = Compute a safe action (**Algorithm 3**)
        Execute safe action a_t , and observe r_t, s_{t+1}
        Compute next observation $\tilde{\mathbf{o}}_{t+1} = f(s_{t+1})$
        Store experience $(\tilde{\mathbf{o}}_t, a_t, r_t, \tilde{\mathbf{o}}_{t+1})$ in RM
        Sample a mini-batch of size M from RM
        Compute $\mathcal{L}(\mathbf{w})$
        Performs gradient descent $\mathbf{w}^+ \leftarrow \mathbf{w} + \alpha \hat{\nabla}_{\mathbf{w}} \mathcal{L}(\mathbf{w})$
      $\mathbf{w}^- = \mathbf{w}^+$ for all $I_i \in \mathcal{I}$
    Every $Target_{update}$ reset $\hat{\mathbf{w}} \leftarrow \mathbf{w}^-$

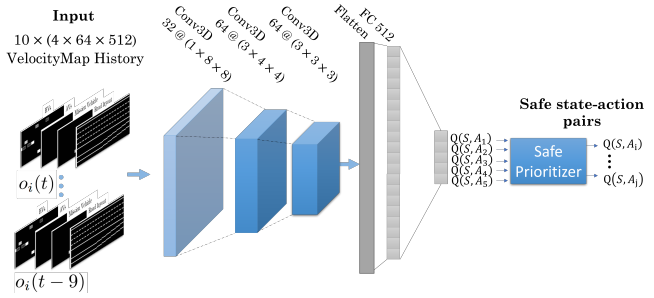

Fig. 3: Deep MARL architecture with the safety prioritizer.

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 $a_t = \max_{a' \in \tilde{\mathcal{A}}_{safe}} Q(\tilde{\mathbf{o}}_t, a'; \mathbf{w})$. 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.

D. Modeling Driver Behaviors

We model the longitudinal movements of HVs using the *Intelligent Driver Model* (IDM) [40], while the lateral actions of HVs are based on the MOBIL model [41]. 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, $\mathrm{a}_n > -b_{\text{safe}}$.

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

$$\mathrm{a}'_{ego} - \mathrm{a}_{ego} + \sin\phi_{ego}\Big((\mathrm{a}'_n - \mathrm{a}_n) + (\mathrm{a}'_o - \mathrm{a}_o)\Big) > \Delta a_{th} \quad (7)$$

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 criterion are mutually satisfied.

The IDM Model determines the longitudinal acceleration of a HV $\dot{v}_{\mathrm{k}}$ as following,

$$\dot{v}_{\mathrm{k}} = \mathrm{a}_{\max}\left[1 - \left(\frac{v_k}{v_{\mathrm{k}}^0}\right)^{\delta} - \left(\frac{d^*(v_k, \Delta v_k)}{d_k}\right)^2\right] \quad (8)$$

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 k^{th} HV, respectively.

The desired minimun gap of the k^{th} HV is given by,

$$d^*(v_k, \Delta v_k) = d_k^0 + v_k T_{\mathrm{k}}^0 + \frac{v_k \Delta v_k}{(2\sqrt{\mathrm{a}_{\max}.\mathrm{a}_{\mathrm{des}}})} \quad (9)$$

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

The typical parameters for MOBIL model are $\sin\phi_e = 0.5$, $\Delta a_{th} = 0.1\frac{m}{s^2}$ and $b_{\text{safe}} = 4\frac{m}{s^2}$. Table I shows typically used parameters of the IDM model [40].

TABLE I: Typical parameters for the IDM model

Parameter	v^0	T^0	$\mathrm{a}_{\max}$	$\mathrm{a}_{\mathrm{des}}$	δ	d^0
Value	30 m/s	1.5 s	1 m/s^2	1.5 m/s^2	4	2 m

Heterogeneous Driver Behaviors Though the parameter from Table I are typical 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 ($\widetilde{\mathcal{H}} \cup \widetilde{\mathcal{I}}$) 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 [27] to classify the behavior (level of aggressiveness) resulted 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 k^{th} vehicle at time t is defined as,

$$\mathcal{C}_C^k[t] = \frac{N-1}{\sum_{v_m \in \mathcal{V}(t)\setminus\{v_k\}} d_t(v_k, v_m)}, \tag{10}$$

where $N = |\widetilde{\mathcal{H}} \cup \widetilde{\mathcal{I}}|$. The more central the vehicle is located, the higher $\mathcal{C}_C^k[t]$ and the closer it is to all other vehicles.

Degree Centrality: the discrete degree centrality of the k^{th} vehicle at time t is defined as,

$$\begin{aligned} \mathcal{C}_D^k[t] &= \left|\{v_m \in \mathcal{N}_k(t)\}\right| + \mathcal{C}_D^k[t-1] \\ \text{such that } &(v_k, v_m) \notin \mathcal{E}(\tau), \tau = 0, \ldots, t-1 \end{aligned} \tag{11}$$

in which $\mathcal{N}_k(t) = \{v_m \in \mathcal{V}(t),\ A_t(k,m) \neq 0, \nu_m \leq \nu_k\}$ represents the set of vehicles in the proximity of the k^{th} vehicle, given that $\nu_m \leq \nu_k$; and ν_m, ν_k denote the velocities of the m^{th} and k^{th} vehicles. The more new vehicles seen by vehicle k that meet this condition, the higher $\mathcal{C}_D^k[t]$.

With the centrality functions we can measure the Style Likelihood Estimate (SLE) for different driver styles [27]. 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,

$$\text{SLE}_l(t) = \left|\frac{\partial \mathcal{C}_C(t)}{\partial t}\right| \quad \text{SLE}_o(t) = \left|\frac{\partial \mathcal{C}_D(t)}{\partial t}\right| \tag{12}$$

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

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 sudden lane changes; while overspeeding the $\mathcal{C}_D(t)$ monotonically increases (higher $\text{SLE}_o(t)$) and during sudden lane changes the slope and the extrema of $\mathcal{C}_C(t)$ will change values. Thus higher values of SLE_{max} are related to increased levels of aggressiveness. Conversely, conservative drivers are not inclined towards those aggressive maneuvers, and the degree 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). Mapping 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 II.

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

E. Computational Details and Hyperparameter

We customized the OpenAI Gym environment in [42] to suit our particular driving scenario and MARL problem. The PyTorch implementation of our architecture on average takes 3.1GB of memory for 4 agents and 18 HVs. Using a GPU NVIDIA Tesla V100. The training process is repeated several times to make sure the experiments converge to a similar policy. The network is trained for $N_{episodes} = 10,000$ taking on average 8 hours and a forward pass during testing requires on average 15ms. We utilize 3,200 GPU-hours for our simulations. Table III lists our simulation and training hyper-parameters.

TABLE II: Estimated simulation parameters that simulate conservative, moderate and aggressive behavior in our scenarios.

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

TABLE III: Simulation and training hyper-parameters.

Parameter	Value	Parameter	Value
N_{episode}	10,000	ϵ decay	Linear
RM buffer size	8,000	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
$\lvert\mathcal{H}\rvert$	18	$\lvert\mathcal{I}\rvert$	4

V. Experimental Results

Controlled Variables We study how the $safe_{th}$, the *level of aggressiveness*, the *traffic scenarios* (f_i) and the *HVs' behaviors* (b_i) impact the performance of AVs. We consider the case in which the mission vehicle (merging/exiting) 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^* been the optimal SVO angle tuned to reach the optimal level of altruism as in [2].

Performance Metrics We measure the performance of our system 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 are used. The altruistic performance gain is measured by computing the difference in the safety/efficiency performance of AV_E and AV_S, as

$$PG_{safety}(\%) = \frac{(AV_E)_{C(\%)} - (AV_S)_{C(\%)}}{N_{Epiosodes}} \tag{13}$$

$$PG_{efficiency}(\%) = \frac{(AV_S)_{DT(m)} - (AV_E)_{DT(m)}}{(AV_E)_{DT(m)}} \tag{14}$$

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,

$$A_{error}(\%) = w_s \times (C(\%)) + w_e \times 100(1 - \frac{DT}{DT_{max}}) \quad (15)$$

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

A. Hypotheses

In this section we examine the following hypotheses

- **H1.** *The higher the level of aggressiveness in a mixed-autonomy scenario, the greater the impact of cooperation. Thus, we expect a higher performance gain (PG) when altruistic AVs face environments with higher level of aggressiveness.*
- **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 ($\mathrm{A_{error}}$).
- **H3.** *With the inclusion of the safety prioritizer, we anticipate improvement in safety and efficiency. We expect that AVs will cause more crashes in the absence of a safety prioritizer ($safe_{th} = 0$).*

B. Analysis and Results

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

1) Sensitivity analyses to HVs behaviors

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 4 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, the higher the impact of cooperation and confirms the **H1**. This is also observed in Figure 5 where the level of aggressiveness is decomposed into lateral and longitudinal aggressiveness. Lateral and longitudinal aggressiveness is varied by changing the MOBIL and IDM parameters (Table II) from aggressive to conservative. Figure 5 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.

2) Domain adaptation of altruistic agents

Following the sensitivity analysis, we investigate the domain adaptation of the AVs to validate the **H2**. Figure 6 shows how the altruistic AVs learn to adapt to different scenarios and behaviors, based on an adaptation score. For the experiments, AV_S are trained in different scenarios $f_i \in \mathcal{F}$ in the presence of HVs with different behaviors $b_j \in \mathcal{B}$ and tested in other scenarios $f_k \in \mathcal{F}$ and behaviors $b_l \in \mathcal{B}$. In our experiments, we consider two case study scenarios $f_m, f_e \in \mathcal{F}$ (merging/exiting) in environments with three different HVs

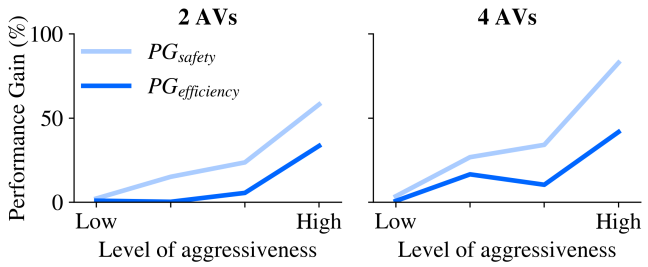

Fig. 4: Sensitivity analyses measured by altruistic performance gain (PG) of AVs, the more aggressiveness of the HVs, the higher the impact/gain of cooperation.

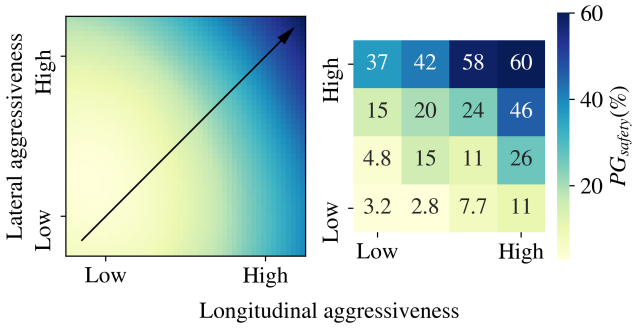

Fig. 5: Lateral and longitudinal sensitivity analyses, the altruistic performance gain (PG) increase in both lateral and longitudinal directions.

behaviors $b_a, b_m, b_c \in \mathcal{B}$ (aggressive, moderate, conservative) see Table II; 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 Figure 6 as an adaptation matrix, showing the $\mathrm{A_{error}}$ for different domains, the $\mathrm{A_{error}}$ is in % and color-map in logarithmic scale to increase the perceived dynamic range for visualization. In our analyses, the weights used for $\mathrm{A_{error}}(\%)$ are $w_s = \frac{2}{3}$ and $w_e = \frac{1}{3}$, which weights the safety performance higher. DT_{max} is computed based on the maximum distance for each situation. Additionally Figure 7 and Figure 8 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 the diagonal as agents trained and tested in the same environment ((f_i, b_j); (f_k, b_l) with $i = k$ and $j = l$) experience during testing similar situations to the ones seen in training. The vehicles trained in the merging environment are able to perform the exiting mission for different behaviors, and vice-versa. It is interesting to notice that when trained in a conservative environment (b_c), the performance when tested in aggressive environments (b_a) is poor. 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 you can

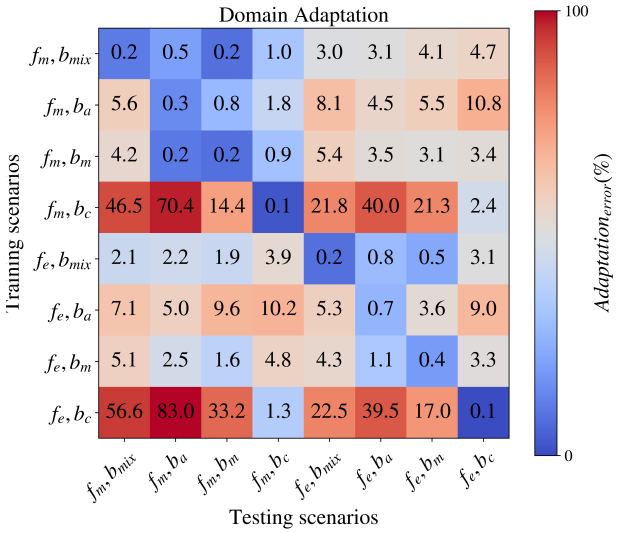

Fig. 6: The domain adaptation matrix with adaptation error ($\mathrm{A_{error}}$) between different traffic scenarios and behaviors. $\mathrm{AV_S}$ are trained (rows of the matrix) in different scenarios $f_i \in \mathcal{F}$ in the presence of HVs with different behaviors $b_j \in \mathcal{B}$ and tested (columns of the matrix) in other scenarios $f_k \in \mathcal{F}$ and behaviors $b_l \in \mathcal{B}$. Each pair (f_i, b_j) is a combination of scenario and behavior. The lower $\mathrm{A_{error}}$ the most suitable the adaptability between those domains.

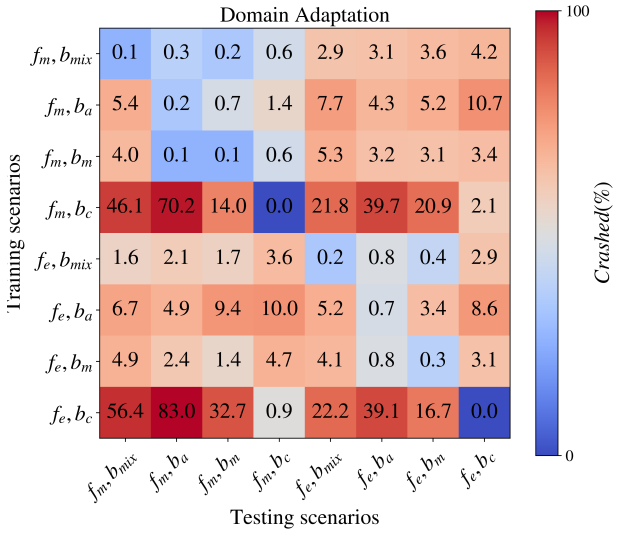

Fig. 7: The domain adaptation matrix with crash percentage ($C(\%)$) between different traffic scenarios and behaviors.The lower $C(\%)$ the most suitable the adaptability in terms of safety (measured by $C(\%)$) between those domains.

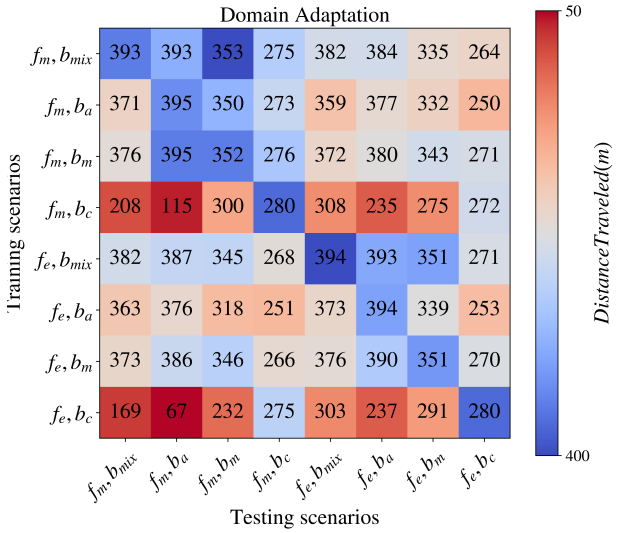

Fig. 8: The domain adaptation matrix with distance traveled ($DT(m)$). Illustrating how the AVs adapt to other situations in terms efficiency (measured by $DT(m)$).

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 $\mathrm{A_{error}}$) 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.

We conclude that the adaptation between the environments is not reciprocal and the selection of the environment and situations should be considered during training, based on the application needs and target situations. The adaptation matrices serve as reference and provide insights on domain adaptation in mixed-autonomy traffic, the matrices present the settings in which altruistic AVs can best learn cooperative policies that are robust to different traffic scenarios and human behaviors.

3) Transfer Learning

Together with domain adaptation, 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 $\mathrm{AV_{merging}}$ (T1), train AVs agents to drive on a highway, and then use that model as the starting point to learn the merging task $\mathrm{AV_{drive-to-merging}}$ (T2), train AVs agents for the exiting task and then use that model as the starting point to learn the merging task $\mathrm{AV_{exiting-to-merging}}$ (T3); and apply the same procedure for the exiting task, learning to exit from scratch $\mathrm{AV_{exiting}}$ (T4), after learned how to drive $\mathrm{AV_{drive-to-exiting}}$ (T5) and after learned how to merge $\mathrm{AV_{merging-to-exiting}}$ (T6). The results of the experiments is presented in Figure 9 and show that our transfer learning approach speeds up the learning process while archiving similar performance as when learning the task from scratch.

4) Safety

Finally, we compared state of the art architectures related to our approach [2], [3], [10], [39] in terms of safety and

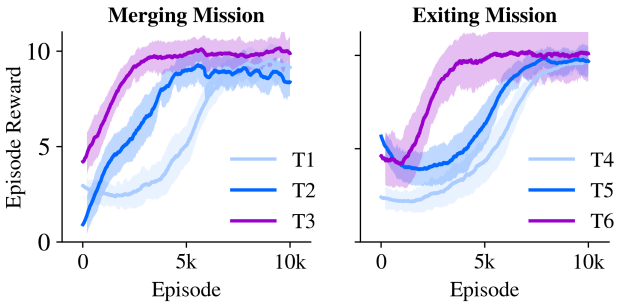

Fig. 9: Transfer learning performance. Showing how policies learned during merging can be transferred to the exiting environment to speed up the learning process while archiving similar performance as when learning the task from scratch.

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 IV our safe altruistic agents consistently outperformed the other approaches, 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 crash in the episodes.

VI. Conclusion and Future Work

We study the problem of multi-agent maneuver-level decision-making in mixed-autonomy environments and investigate how AVs can learn cooperative policies that are robust to different scenarios and driver behaviors safely. Our 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 safe altruistic AVs learn to coordinate and influence the behavior of HVs with socially advantageous results in diverse situations.

Limitations and Future Work. While we explored different aspects of social navigation in various environments and in the presence of diverse HVs behaviors, the HV models are not learned from real human drivers' data and the traffic scenarios are limited to merging and exiting. Nevertheless, we speculate that our approach could be effective in realistic traffic situations by utilizing and learning from real human data and traffic scenarios. Additionally, to be used in the real world, extra emphasis is needed on safety.

In future work, we plan to investigate more sophisticated architecture and state representations, 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 and diverse traffic agents such as bicycles or pedestrians. Despite the limitations, we are thrilled to see safe and robust social AVs on the road that learn from experience. We also anticipate applications of these ideas beyond driving, to general MA humans-robot interactions when agents influence humans and cooperate safely for a socially advantageous outcome.

References

[1] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus, "Social behavior for autonomous vehicles," *Proceedings of the National Academy of Sciences*, vol. 116, no. 50, pp. 24 972–24 978, 2019.

[2] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "Social coordination and altruism in autonomous driving," *arXiv preprint arXiv:2107.00200*, 2021.

[3] ——, "Cooperative autonomous vehicles that sympathize with human drivers," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021.

[4] D. Sadigh, N. Landolfi, S. S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state," *Autonomous Robots*, vol. 42, no. 7, pp. 1405–1426, 2018.

[5] F. Sagberg, Selpi, G. F. Bianchi Piccinini, and J. Engström, "A review of research on driving styles and road safety," *Human factors*, vol. 57, no. 7, pp. 1248–1275, 2015.

[6] P. B. Harris, J. M. Houston, J. A. Vazquez, J. A. Smither, A. Harms, J. A. Dahlke, and D. A. Sachau, "The prosocial and aggressive driving inventory (padi): A self-report measure of safe and unsafe driving behaviors," *Accident Analysis & Prevention*, vol. 72, pp. 1–8, 2014.

[7] E. F. Vallières, R. J. Vallerand, J. Bergeron, and P. McDuff, "Intentionality, anger, coping, and ego defensiveness in reactive aggressive driving," *Journal of Applied Social Psychology*, vol. 44, no. 5, pp. 354–363, 2014.

[8] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for autonomous cars that leverage effects on human actions." in *Robotics: Science and Systems*, vol. 2. Ann Arbor, MI, USA, 2016.

[9] C. Wu, A. Kreidieh, E. Vinitsky, and A. M. Bayen, "Emergent behaviors in mixed-autonomy traffic," in *Conference on Robot Learning*. PMLR, 2017, pp. 398–407.

[10] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "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.

[11] Z. Li, U. Kalabić, and T. Chu, "Safe reinforcement learning: Learning with supervision using a constraint-admissible set," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 6390–6395.

[12] A. Cosgun, L. Ma, J. Chiu, J. Huang, M. Demir, A. M. Anon, T. Lian, H. Tafish, and S. Al-Stouhi, "Towards full automated drive in urban environments: A demonstration in gomentum station, california," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1811–1818.

[13] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, "A survey of learning in multiagent environments: Dealing with non-stationarity," *arXiv preprint arXiv:1707.09183*, 2017.

[14] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1086–1095, 2019.

[15] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, "Reinforcement learning-based multi-agent system for network traffic signal control," *IET Intelligent Transport Systems*, vol. 4, no. 2, pp. 128–135, 2010.

[16] A. OroojlooyJadid and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," *arXiv preprint arXiv:1908.03963*, 2019.

[17] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[18] A. Xie, D. Losey, R. Tolsma, C. Finn, and D. Sadigh, "Learning latent representations to influence multi-agent interaction," in *Proceedings of the 4th Conference on Robot Learning (CoRL)*, November 2020.

[19] A. Shih, A. Sawhney, J. Kondic, S. Ermon, and D. Sadigh, "On the critical role of conventions in adaptive human-ai collaboration," in *9th International Conference on Learning Representations (ICLR)*, 2021.

[20] E. Vinitsky, R. Köster, J. P. Agapiou, E. Duéñez-Guzmán, A. S. Vezhnevets, and J. Z. Leibo, "A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings," *arXiv preprint arXiv:2106.09012*, 2021.

[21] A. Pokle, R. Martín-Martín, P. Goebel, V. Chow, H. M. Ewald, J. Yang, Z. Wang, A. Sadeghian, D. Sadigh, S. Savarese *et al.*, "Deep local trajectory replanning and control for robot navigation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5815–5822.

TABLE IV: Architectures' performance comparison. Our safe altruistic AVs outperformed the others approaches.

	Aggressive HVs			Moderate HVs			Conservative HVs		
Approaches	C (%)	MF (%)	DT (m)	C (%)	MF (%)	DT (m)	C (%)	MF (%)	DT (m)
Conv2D+DQN [39]	31.2	28.9	316	25.4	20.3	302	14.0	7.9	274
Toghi *et al.* [2]	21.3	16.4	339	12.7	10.1	333	1.6	0.6	269
Conv3D+A2C [10]	14.8	12.6	341	9.4	8.8	328	1.1	0.1	267
Conv3D+DQN [3]	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*

[22] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

[23] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2681–2690.

[24] K. Brown, K. Driggs-Campbell, and M. J. Kochenderfer, "A taxonomy and review of algorithms for modeling and predicting human driver behavior. arxiv e-prints, article," *arXiv preprint arXiv:2006.08832*, 2020.

[25] B. Ivanovic, E. Schmerling, K. Leung, and M. Pavone, "Generative modeling of multimodal multi-human behavior. in 2018 ieee," in *RSJ International Conference on Intelligent Robots and Systems*, pp. 3088–3095.

[26] K. H. Beck, B. Ali, and S. B. Daughters, "Distress tolerance as a predictor of risky and aggressive driving," *Traffic injury prevention*, vol. 15, no. 4, pp. 349–354, 2014.

[27] R. Chandra, U. Bhattacharya, T. Mittal, A. Bera, and D. Manocha, "Cmetric: A driving behavior measure using centrality functions," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2035–2042.

[28] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, "Driving style analysis using data mining techniques," *International Journal of Computers Communications & Control*, vol. 5, no. 5, pp. 654–663, 2010.

[29] M. Kuderer, S. Gulati, and W. Burgard, "Learning driving styles for autonomous vehicles from demonstration," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2641–2646.

[30] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," *Advances in neural information processing systems*, vol. 29, pp. 3909–3917, 2016.

[31] C. Wu, A. M. Bayen, and A. Mehta, "Stabilizing traffic with autonomous vehicles," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6012–6018.

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

[33] D. A. Lazar, E. Bıyık, D. Sadigh, and R. Pedarsani, "Learning how to dynamically route autonomous vehicles on shared roads," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103258, 2021.

[34] J. Wang, Q. Zhang, D. Zhao, and Y. Chen, "Lane change decision-making through deep reinforcement learning with rule-based constraints," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.

[35] S. Nageshrao, H. E. Tseng, and D. Filev, "Autonomous highway driving using deep reinforcement learning," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 2326–2331.

[36] C.-J. Hoel, K. Driggs-Campbell, K. Wolff, L. Laine, and M. J. Kochenderfer, "Combining planning and deep reinforcement learning in tactical decision making for autonomous driving," *IEEE transactions on intelligent vehicles*, vol. 5, no. 2, pp. 294–305, 2019.

[37] A. Mohammadhasani, H. Mehrivash, A. Lynch, and Z. Shu, "Reinforcement learning based safe decision making for highway autonomous driving," *arXiv preprint arXiv:2105.06517*, 2021.

[38] D. Chen, Z. Li, Y. Wang, L. Jiang, and Y. Wang, "Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic," *arXiv preprint arXiv:2105.05701*, 2021.

[39] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[40] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[41] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.

[42] E. Leurent, Y. Blanco, D. Efimov, and O.-A. Maillard, "Approximate robust control of uncertain dynamical systems," *arXiv preprint arXiv:1903.00220*, 2019.

Rodolfo Valiente is a Ph.D. candidate in Computer Engineering at the University of Central Florida (UCF). His research interests include connected autonomous vehicles (CAVs), reinforcement learning, computer vision, and deep learning with a focus on the autonomous driving problem. He received a M.Sc. degree from the University of Sao Paulo (USP) and his B.Sc. degree from the Technological University Jose Antonio Echeverria.

Behrad Toghi is a Ph.D. candidate at the University of Central Florida. He received the B.Sc. degree in electrical engineering from Sharif University of Technology in 2016 and has worked as a research intern at Honda Research Institute, Mercedes-Benz R&D North America and Ford Motor Company R&D between 2018 and 2021. His work is mainly in the domains of prediction & behavior planning for autonomous driving.

Ramtin Pedarsani is an Assistant Professor in the ECE Department at the University of California, Santa Barbara. He received the B.Sc. degree in electrical engineering from the University of Tehran in 2009, the M.Sc. degree in communication systems from the Swiss Federal Institute of Technology (EPFL) in 2011, and his Ph.D. from the University of California, Berkeley, in 2015. His research interests include networks, game theory, machine learning, and transportation systems.

Yaser P. Fallah is an Associate Professor in the ECE Department at the University of Central Florida. He received the Ph.D. degree from the University of British Columbia, Vancouver, BC, Canada, in 2007. From 2008 to 2011, he was a Research Scientist with the Institute of Transportation Studies, University of California Berkeley, Berkeley, CA, USA. His research, sponsored by industry, USDoT, and NSF, is focused on intelligent transportation systems and automated and networked vehicle safety systems.