Social Coordination and Altruism in Autonomous Driving

Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, *Senior Member, IEEE*, and Yaser P. Fallah

***Abstract*— Despite the advances in the autonomous driving domain, autonomous vehicles (AVs) are still inefficient and limited in terms of cooperating with each other or coordinating with vehicles operated by humans. A group of autonomous and human-driven vehicles (HVs) which work together to optimize an altruistic social utility can co-exist seamlessly and assure safety and efficiency on the road. Achieving this mission without explicit coordination among agents is challenging, mainly due to the difficulty of predicting the behavior of humans with heterogeneous preferences in mixed-autonomy environments. Formally, we model an AV's maneuver planning in mixed-autonomy traffic as a partially-observable stochastic game and attempt to derive optimal policies that lead to socially-desirable outcomes using a multi-agent reinforcement learning framework (MARL), and propose a semi-sequential multi-agent training and policy dissemination algorithm for our MARL problem. We introduce a quantitative representation of the AVs' social preferences and design a distributed reward structure that induces altruism into their decision-making process. Altruistic AVs are able to form alliances, guide the traffic, and affect the behavior of the HVs to handle competitive driving scenarios. We compare egoistic AVs to our altruistic autonomous agents in a highway merging setting and demonstrate the emerging behaviors that lead to improvement in the number of successful merges and the overall traffic flow and safety.**

***Index Terms*— Cooperative driving, social navigation, mixed-autonomy traffic, multi-agent reinforcement learning.**

I. Introduction

CONNECTED and automated vehicles (CAVs) pursue a mission to enhance driving safety and reliability by bringing automation and intelligence into vehicles, which lessens the inherent human limitations such as range of vision, reaction time, and distraction. Adding the communication component to intelligent vehicles further improves their ability to perceive their surroundings and creates an opportunity for mass coordination and cooperative decision-making. This inter-agent coordination is particularly important as the full potential of CAVs does not lie in operating a single vehicle on an empty road but rather from their seamless co-existence with other autonomous and human-driven vehicles (HVs). Hence, we narrow the focus of this work to studying the decision-making problem in the presence of multiple autonomous agents and human drivers, i.e. a mixed-autonomy multi-agent environment.

Leveraging vehicle-to-vehicle (V2V) communication, decision-making in a purely-autonomous environment can be simplified into a centralized control problem with essentially one agent. However, the presence of HVs makes inter-agent coordination more challenging as they cannot explicitly communicate to coordinate with AVs in real-time. In order to make safe and socially-desirable decisions in the presence of humans, current solutions on social navigation for AVs mainly rely on learned or hand-coded models that predict the behavior of human drivers [1], [2]. We identify two key shortcomings in the existing schemes. First, the fidelity of the human models that are derived in the absence of autonomous agents is questionable in mixed-autonomy settings as human drivers tend to act differently when around AVs [3]. Second, single-agent solutions do not fully exploit the potential of CAVs in constituting a mass intelligence, forming alliances, and performing coordinated multi-agent maneuvers.

We study the mixed-autonomy decision-making problem from a multi-agent point of view, as opposed to the previous individual perspectives. Our key insight is that incentivizing AVs on adopting an *altruistic behavior* and accounting for the interest of other vehicles, allows them to see the big picture and find solutions that are optimal for the group in the longer term. In addition to the potential safety and efficiency benefits of altruistic decision-making, altruism leads to circumstances where no vehicle has superiority over the others, creating more societally beneficial outcomes [4]. To elaborate, Figure 1(a) shows that a group of AVs can guide the behavior of human drivers to improve safety and efficiency, Figures 1(b) and 1(c) illustrate examples of how AVs can work together to achieve a social goal that benefits another HV or AV.

We focus our work on inherently competitive driving scenarios, such as the examples illustrated in Figure 1, where safe and efficient traffic flow necessarily requires coordination among autonomous agents and egoistic behavior most likely compromises traffic safety and efficiency. We build on our

Manuscript received 21 February 2022; revised 1 August 2022; accepted 13 September 2022. Date of publication 29 September 2022; date of current version 5 December 2022. This work was supported in part by the National Science Foundation under Grant CNS-1932037, Grant 1953032, and Grant 1952920. The Associate Editor for this article was C. Lv. *(Behrad Toghi and Rodolfo Valiente contributed equally to this work.) (Corresponding author: Rodolfo Valiente.)*

Behrad Toghi, Rodolfo Valiente, and Yaser P. Fallah are with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: rvalienter90@knights.ucf.edu).

Dorsa Sadigh is with the Department of Electrical Engineering and the Department of Computer Science, Stanford University, Stanford, CA 94305 USA.

Ramtin Pedarsani is with the Department of Electrical and Computer Engineering, UC Santa Barbara, Santa Barbara, CA 93106 USA.

Digital Object Identifier 10.1109/TITS.2022.3207872

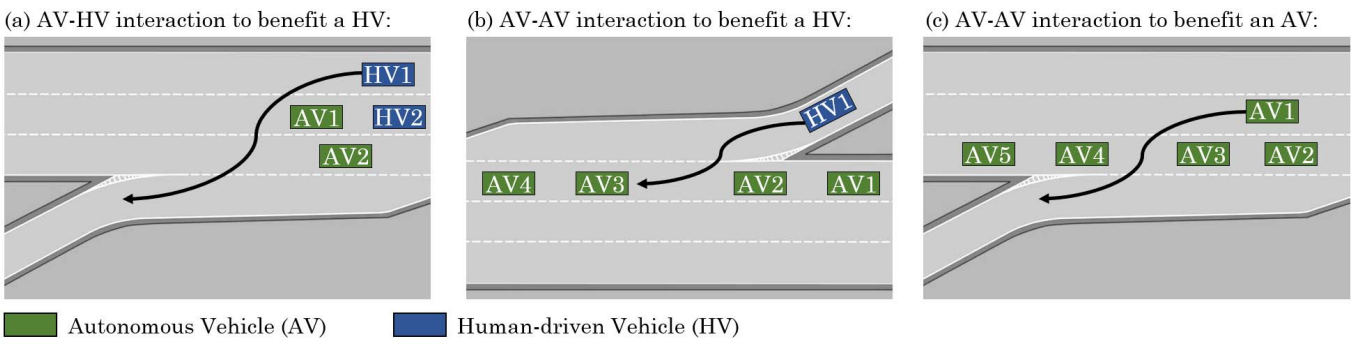


Fig. 1. **(a) AV-HV interaction to benefit another HV:** Altruistic agents have the opportunity to form alliances and guide the behavior of HVs in order to improve the traffic flow and avoid hazardous situations. AV1 & AV2 can build a formation to slow down HV2 and open up a pathway for HV1, enabling it to trust the AVs, change lanes, and navigate towards the exit ramp. **(b) AV-AV interaction to benefit another HV:** HV1 is intended to merge into the highway. Egoistic AVs ignore the merging vehicle and do not open up space for it which can potentially lead to hazardous scenarios, whereas if they show sympathy for the merging HV, they can compromise on their own interest in order to create a safe path for HV1 to merge into the highway. **(c) AV-AV interaction to benefit another AV:** AV1 attempts to exit the highway. If AV2-AV5 act egoistically, AV1 might miss the exit and not be able to follow its planned mission. However, if AV2-AV5 take into account the interest of AV1 and act altruistically, they can open up space in the platoon, by AV2 & AV3 decelerating and AV4 & AV5 accelerating, to enable a safe exit for AV1.

prior work in [5] and [6] and proposed a novel semi-sequential multi-agent training and policy dissemination algorithm to alleviate the non-stationary problem. Additionally, we use a method for scoring the entries in the experience replay buffer that improves sample efficiency and speeds up the learning process. Furthermore, we emphasize the importance of finding the optimal social value orientation and in contrast to the other works, formulate it as a convex optimization problem. We formalize the mixed-autonomy driving problem as a partially observable stochastic game (POSG) and derive optimal policies using deep multi-agent reinforcement learning (MARL). With our solution, altruistic autonomous agents not only learn to drive safely but also master inter-agent coordination and social navigation. Our main contributions are as follows:

- We propose a MARL framework to train altruistic agents using a decentralized social reward signal. These agents are able to drive safely on the highway and coordinate with each other in the presence of human drivers.
- We proposed a novel semi-sequential multi-agent training and policy dissemination algorithm for our MARL problem and utilized a network architecture that allows our agents to implicitly learn from experience, without the need for an explicit behavioral model of human drivers.
- In contrast with the existing solutions, we formulate the problem of finding the optimal social value orientation angle as a convex optimization objective. We show that an optimal value for the level of altruism exists and when chosen properly between being absolutely selfless or selfish, despite some agents' compromise on their local utility, the overall traffic safety, and flow improve for the group of vehicles.

II. Related Work

This section presents a short literature review on the main topics that are closely related to our problem, namely core MARL solutions, cooperative algorithms, human behavior modeling, and navigation in the presence of humans.

A. Multi-Agent Reinforcement Learning

Early solutions for multi-agent value-learning algorithms assume independently trained agents and are proved to perform poorly [7]. To alleviate this problem, a learning rule is presented by Foerster *et al.* that relies on an additional term to take into account the effect of other agents' evolution during the training. They have also attempted to leverage a multi-agent derivation of importance sampling and remove outdated samples from the experience replay buffer [8] to make it effective for multi-agent settings. Xie *et al.* employs latent representations of partner strategies to address this problem and enable a more scalable partner modeling [9]. Shih *et al.* further considers the effects of repeated interactions on partner modeling and develop a modular approach that separates rule-dependent representations from partner-dependent conventions [10].

Foerster *et al.* proposed the counterfactual multi-agent (COMA) algorithm that is expected to address the credit assignment problem in multi-agent environments [11]. COMA algorithm utilizes the set of joint actions of all agents as well as the full state of the world during the training. In contrast, we assume partial observability and a decentralized reward function during both training and execution. More application-oriented related works include the centralized multi-agent solutions proposed by Gupta *et al.* [12]. More recently, Wang *et al.* proposed a gifting approach that enables the emergence of prosocial behaviors in general-sum coordination games [13]. Importantly, in contrast with our approach, the existing literature on multi-agent systems relies on assumptions on the social preference of agents [14], [15].

B. Human Behavior Modeling

Driving styles of human drivers can be learned either from demonstration through inverse RL, as proposed by

Kuderer *et al.* , or employing statistical models such as Gaussian and Dirichlet processes [16], [17]. Kuefler *et al.* adopt a novel approach and apply generative adversarial networks to imitate the behavior of a human driver [18]. Schmerling *et al.* study the scenarios with inherent multimodal uncertainty, such as our driving, and leverage conditional variational autoencoders (CVAEs) to condition the policy on the present interaction history [19]. Recent data-driven approaches have shown achievements in classifying human driving maneuvers [20] and predicting human trajectories to enable fully-autonomous navigation of a robot in human-dense environments [21]. In contrast with works in the broad literature on human behavior modeling that take a game-theoretic or optimization-based approach, we rely on implicitly learning from interaction data within our MARL platform.

C. Social Value Orientation

Schwarting *et al.* initially proposed the idea of leverag-ing SVO in the form of an angular representation to study and govern the social behavior of autonomous vehicles in a game-theoretic setup, which contrasts with our reinforcement learning framework [4]. Crosato *et al.* address an important problem, i.e., the freezing robot problem, that is a consequence of agents overestimating risks created by humans. Our work is different in the sense that we address the broader problem of training altruistic agents which can impact and govern the road behavior of the vehicles around them [22]. Van Vugt *et al.* also employed the concept of SVO to address some of the ethical and societal questions around AVs [23]. Geary *et al.* introduce the notion of "Information Sufficiency" and leverage that to define a reward function that enables AV to choose actions with altruistic nature. The authors formulate the mixed-autonomy driving problem as a Stackelberg Game while we view it as a Markov Game and solve it using a Muli-agent RL framework [24].

D. Robot Navigation

Alahi *et al.* introduced the Social LSTM framework which leverages recurrent neural networks to extract the tempo-ral information from the trajectory of pedestrians in large crowds [1]. Tsoi *et al.* present their high-fidelity simulation platform, SEAN, to accelerate the research on social robot navigation [25]. Vazquez *et al.* study the social interactions in a human-robot role-playing game and expand their observations to study the spatial behavior of a group of robots. More recent works in social navigation have revealed the potential for collaborative planning and interaction with humans. Examples include but are not limited to works by Trautman *et al.* and Nikolaidis *et al.* where a mutual reward function is optimized in order to enable joint trajectory planning for humans and robots [26], [27].

E. Mixed-Autonomy Traffic Networks

Lazar *et al.* take a more abstract and traffic-level perspective to study the emergent behaviors in mixed-autonomy environments using model-free RL solutions [28]. Wu *et al.* explore the idea of stabilizing the traffic flow that is guided by autonomous vehicles as well as the emergent behaviors in a mixed AV-HV setting [29], [30]. Vinitsky *et al.* present a benchmark for traffic control based on RL in mixed-autonomy traffic [31]. Biyik *et al.* formalize the effects of altruistic driving in mixed-autonomy at a road level and present a formal model of road congestion that can be used for optimal routing in road networks [32].

III. Preliminaries

In this section, we provide the preliminary concepts that are essential in the following section and introduce our formal notation.

A. Partially-Observable Stochastic Games

Decision-making process in a finite set of autonomous agents I with partial observability in stochastic environments can be formalized as a partially-observable stochastic game (POSG) defined by the tuple $M_G := (I, S, [A_i], [O_i], T, [R_i])$ for $i = 1, \ldots, N$. At a given time, each agent receives a local observation $\mathbf{o}_i : S \rightarrow O_i$ that is correlated with the underlying state of the environment $s \in S$ and takes an action from the action space $a \in A$. Consequently, the environment evolves to a new state s_i^0 with probability $T = \Pr(s^0 | s, a) : S \times A_1 \times \ldots \times A_N \rightarrow S$ and the agent receives a decentralized reward $R_i : S \times A_i \rightarrow \mathbb{R}$. The probability distribution over actions at a given state is known as the stochastic policy $\pi_i : O_i \times A_i \rightarrow [0, 1]$. The goal is to derive a distribution that maximizes the discounted sum of future rewards over an infinite time horizon, i.e., an optimal policy $\pi^* : S \rightarrow A$,

$$\pi^* := \arg\max_{\pi} \mathbb{E} \sum_{k=0}^{\infty} \gamma^k R\left(s_k, \pi(s_k)\right) \tag{1}$$

in which, $\gamma \in [0, 1)$ is the discount factor. The optimal policy maximizes the state-action value function, i.e., $\pi^*(s) = \arg\max_a Q^*(s, a)$, where

$$Q\pi(s, a) := \mathbb{E} \sum_{k=0}^{\infty} \gamma^k R\left(s_k, \pi(s_k)\right) \mid s_0 = s, a_0 = a \tag{2}$$

and the optimal state-action value function can then be derived using the Bellman optimality equation,

$$Q^*(s, a) = \mathbb{E}_{s^0 \sim P(. \mid s,a)} R(s, a) + \max_{a^0} \gamma\, Q^*(s^0, a^0) \tag{3}$$

B. Solving POSGs With Unknown Dynamics

The dynamics of the environment and reward function are usually stochastic and not fully known in real-world problems. Reinforcement learning (RL) provides a possibility to solve POSGs with unknown rewards and state transition functions through continuous interaction with the environment. RL algorithms such as off-policy temporal difference learning enable

agents to update the value function from such interactions with the environment,

$$Q_{k+1}(s,a) - Q_k(s,a) = \alpha_k \left[R(s,\pi(s)) + \gamma \max Q_k(s^0,a^0) - Q_k(s,a) \right], \quad (4)$$

where α_k is the learning rate at the kth iteration.

C. Deep Q-Networks

Parameterizing the state-action value function using a function approximator, i.e., $\tilde{Q}(.;\mathbf{w}) \approx Q(\cdot)$, results in more generalizable policies that can scale to larger state-spaces. Parameters $\mathbf{w}$ can be learned through mini-batch gradient descent steps,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k \hat{\nabla}_{\mathbf{w}} L(\mathbf{w}_k) \quad (5)$$

where, the $\hat{\nabla}_{\mathbf{w}}$ operator estimates the gradient at $\mathbf{w}_k$. Deep neural networks are widely used as function approximators and are also applicable to the Q-learning algorithm [33]. A deep Q-network (DQN) builds up on two major ideas, namely using two separate networks during training and employing an experience replay buffer to decorrelate the training samples. The former is done to stabilize the training process by updating the greedy network at each training iteration to compute the optimal Q-value and using another less-frequently updated target network. The loss function in Eq. (5) can be written as

$$L(\mathbf{w}_k) = \mathbb{E}\left[R + \gamma \max_{a'} Q^{\hat{}}(s^0,a^0;\hat{\mathbf{w}}) - Q^{\hat{}}(s,a;\mathbf{w}) \right]^2 \quad (6)$$

where $\hat{\mathbf{w}}$ is the target network that periodically gets updated during the training. Additionally, the DQN algorithm draws batches of training data (s, a, R, s^0) from an experience replay buffer in order to decorrelate the training samples in Eq. (5) that are generated from simulation or real-world experience and thus naturally have temporal dependencies. This process is challenging in MARL since, $\Pr(s^0|s,a,\pi_1,\ldots,\pi_n) = \Pr(s^0|s,a,\pi_1^0,\ldots,\pi_n^0)$ if any $\pi_i = \pi_i^0$. In other words, the environment becomes non-stationary when multiple agents are evolving concurrently. We will further discuss this issue and provide a solution to stabilize the multi-agent learning process in Section V-D.

D. V2V Networks

We are interested in a multi-agent setting where agents have no information about others' actions and cannot explicitly coordinate. Instead, the decentralized coordination among agents is expected to arise from the social reward signal. We extend the earlier introduced concepts to a coordinated POSG defined as $(I, S, [A_i], [O_i], T, [R_i], G)$, where $G = (I, E)$ is a stochastic, time-varying, undirected graph that encompasses the V2V communication among agents in the environment E. The communicated information can be as simple as kinematics information, e.g., speed, location, heading, or more bandwidth-intensive forms of sensory data, e.g., camera and LiDAR. Leveraging this shared situational awareness, agents can extend their range of perception and overcome obstacles and line-of-sight visibility limitations [34], [35]. An agent's local observation $\tilde{\mathbf{o}}_i \subseteq O_i$ is created using the shared situational awareness and clearly depends on G which incorporates the flow of information among agents. We utilize the network analysis from [36] to model the V2V communication on a high-density highway.

IV. Problem Statement

We investigate the maneuver-level decision-making problem for AVs to explore behaviors that can lead to socially-desirable outcomes. We are interested in the question of how autonomous agents can be trained from scratch to perform an individual task such as driving safely on a road, while considering the social aspects of their mission, i.e., optimizing for a social utility that also accounts for the interest of other vehicles around them. Figure 1 helps us to build more intuition on the topic by depicting instances of driving scenarios in which altruism leads to socially-valuable outcomes and clearly overcomes the limitations of egoistic and single-agent planning. Each example in Figure 1 provides an example of altruistic inter-agent coordination settings that can benefit both HVs and AVs. It is clear that in some instances, altruistic AVs have to compromise on their individual utility, e.g., by slowing down, in order to increase the group's overall utility. The balance between an AV's selflessness and selfishness is the key to reaching efficient and safe traffic flow. In [5] and [6] we show that tuning the level of altruism in AVs leads to different emerging behaviors and affects the traffic flow and driving safety. In this work, we further explore that finding and formulate the problem as a convex optimization objective, to obtain an optimal social value orientation angle. Thus, we continue this section by providing a quantitative representation of an agent's level of altruism and formally defining our case study scenario, before presenting our proposed solution in the next section.

A. Quantifying Social Value Orientation

In order to formally study the social dilemmas between humans and autonomous agents in heterogeneous environments, it is crucial to quantify the social preference of an individual, e.g., whether they will defect or cooperate in a given situation such as opening a gap in our highway merging example. The degree of an agent's egoism or altruism with regard to its counterparts is defined as *Social Value Orientation (SVO)*, a widely used notion in the social psychology literature that has been recently adopted in robotics research. Specifically, we borrow the angular annotation for SVO as defined by Liebrand *et al.* [37]. The SVO angular preference φ, quantifies how an agent weights its own reward against the reward of others. An agent's total utility R_i can then be written as,

$$R_i = r_i \cos\varphi_i + r_i^- \sin\varphi_i \quad (7)$$

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

Fig. 2. SVO angular phase φ quantifies an agent's level of altruism. The figure is based on the empirical data collected from humans by Garapin *et al.* [38]. The diameter of the circles shows the size of the human population that holds the corresponding SVO.

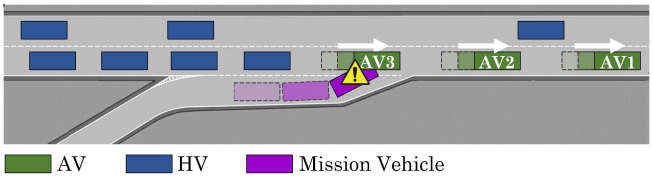


Fig. 3. Case study: a mission vehicle that can be human-driven or autonomous attempts to merge into the mixed group of AVs and HVs.

general is a function $f(\cdot)$ of their individual utilities,

$$r_i^- = f(r_j), \quad \text{where } j = i \tag{8}$$

Autonomous agents require an understanding of human drivers' social preferences and their willingness to coordinate. However, it is well-established in the behavioral decision theory that humans are heterogeneous in SVO and thus their preference is rather ambiguous and unclear [39]. Current works on social navigation for AVs often make restrictive assumptions about human drivers' social preference and compliance [2], whereas Figure 2 indicates a spectrum of altruism among humans with heterogeneous social value orientations. Thus, due to the large spectrum of altruistic behavior observed by humans, our insight is to rely on autonomous cars instead to guide the overall system toward more socially desirable objectives. Specifically, we plan to find policies for AVs that improve the utility of the group as a whole through emerging alliances and more importantly, affecting the behavior of human drivers. In our particular driving example, the desired social outcome is achieving seamless and safe highway merging while maximizing the distance traveled by all vehicles and avoiding collisions.

B. Formalism

We choose a highway merging scenario with a mixed group of AVs and HVs as our base experiment scenario, as illustrated in Figure 3. A merging vehicle, which can be either HV or AV, approaches the highway on the merging ramp and faces a mixed platoon of vehicles that are cruising on the highway. This configuration contains a group of AVs that hold the same SVO, as well as a group of HVs that are heterogeneous in their SVO, hence it is unclear if they are allies or foes. In this setting, it is obvious that the individual interest of the merging vehicle, i.e., seamless merging into the highway, does not align with that of the cruising vehicles, i.e., cruising with optimal speed and energy consumption. We design our case study scenario in a way that safe and seamless merging necessarily requires all AVs to work together and none of them alone can enable the merging of the mission vehicle without the cooperation of the others. Formally, the road section shown in Figure 3 is shared by a set of AVs I that are connected together via V2V communication and governed by a decentralized stochastic policy, a set of HVs V operated by humans with heterogeneous and unknown SVOs, and a human-driven or autonomous *mission vehicle* $M \notin I \notin V$ that attempts to merge into the highway.

A human driver's perception is often limited by their range of vision, occlusion, and obstacles. In contrast, CAVs share their observations to overcome these limitations. Each CAVs has a unique local observation $\tilde{\mathbf{o}}_i([\mathbf{o}_i]; G)$ that is constructed using its own local observation, as well as the local obser-vations it receives from the neighboring CAVs. As mentioned before, graph G grasps this inter-agent communication. There-fore, an observer AV can detect a subset of other AVs, $I \notin I$, and a subset of HVs $V \notin V$. As we elaborated before, our aim is to find a decentralized control scheme that can induce altruism in the behavior of AVs. Hence, each AV must use its local observation $\mathbf{o}_i$ to make independent decisions that optimize its utility. The value of the agent's altruism, i.e., the SVO angular phase φ, determines the social implications of an agent's local actions. To summarize, we state our problem as *deriving a utility function that enables the AVs to handle competitive driving scenarios, such as those illustrated in Figure 1, and lead them into socially-desirable outcomes that improve traffic safety and efficiency for the group of vehicles*.

V. Sympathetic Cooperative Driving Framework

In their recent work, Silver *et al.* explained how artificial intelligence agents can learn complex tasks through experience and maximizing a generic reward function, rather than requiring task-specific specialized problem formulations [40]. Inspired by this approach to solving decision-making problems, rather than breaking down our problem into *learning how to drive* and *learning social coordination*, we train our autonomous agents from scratch using a decentralized reward structure and expect them to master the basics of highway driving, e.g., avoiding collisions and unnecessary lane change or acceleration, while learning inter-agent coordination to eventually achieve the goal of enabling a safe and seamless merging. To reiterate our goal, we seek a decentralized solution that enables autonomous agents to make independent socially-desirable decisions, with no explicit coordination or sharing of their decisions and future actions. In the rest of this section, we define the action and observation space in the POSG framework of Section III and introduce the notions of sympathy and cooperation that are essential for structuring the reward function.

A. Action and Observation Spaces

We employ a numeric representation for an agent's observation that embeds the kinematics of the neighboring vehicles.

Additionally, we integrate the history of vehicles' last h meta-actions to extract temporal information and their past trajectories. An ego vehicle $I_i \in I$ observes a set of HVs and AVs in its perception range. The *Kinematic* observation includes the relative Frenet coordinates of the closest $|I \cap V| + 1$ vehicles in addition to the absolute Frenet coordinates of the ego vehicle. Formally, agent I_i receives a local observation $\tilde{\mathbf{o}}_i \in O_i$,

$$\tilde{\mathbf{o}}_i = \left[o_i, o_m, o_{i+1}, \ldots, o_{i+|I \cap V|}\right]^\top \tag{9}$$

Each row of the local observation matrix $\mathbf{o}^{(i)}$ is defined as,

$$o_j = \left[p_j, l_j, d_j, \mathrm{d}l_j/\mathrm{d}t, \mathrm{d}d_j/\mathrm{d}t, \cos\rho_j, \sin\rho_j, \lambda_j, \overline{\mathrm{H}}_j^A\right] \tag{10}$$

in which, l_j and d_j are the longitudinal and lateral Frenet coordinates of the jth vehicle, respectively. Vehicle's yaw angle is denoted by ρ and the autonomy flag is $\lambda_j = 0$ if $I_j \in V$ and $\lambda_j = 1$ otherwise. In case the total number of observed vehicles is smaller than the set size of the observation matrix $\mathbf{o}$, the remaining rows are filled with zeros with $p_j = 0$.

$\overline{\mathrm{H}}_j^A$ is the unrolled numeric representation of the action history array $\mathbf{H}_j^A$ that contains the last h meta-actions taken by I_j and is defined as,

$$\mathbf{H}_j^A(t) = \left[a_j(t-1), \ldots, a_j(t-h)\right] \tag{11}$$

Our interest is in maneuver-level decision-making for autonomous vehicles. Thus, we define the action space A as the set of abstract meta-actions $A_i = [\texttt{Lane Left}, \texttt{Idle}, \texttt{Lane Right}, \texttt{Accelerate}, \texttt{Decelerate}]$. These meta-actions are then translated into admissible tra-jectories and low-level control signals that eventually govern the movement of the vehicle. The implementation details of how meta-actions render into steering and acceleration signals are discussed in Section VI. Additionally, the discrete meta-actions defined above must be translated into numeric values in Eq. (11). We experiment with three encodings and choose the one that leads to the best performance after training:

- *Binary*: A one-hot encoding with 5 bits for $a_i \in A_i$.
- *Discrete*: An integer in $(0, 5]$ for $a_i \in A_i$.
- *Frenet*: Two integers in $[-1, 1]$ for lateral and longitudinal actions.

B. Disentangling Sympathy and Cooperation

Inter-agent relations in our mixed-autonomy problem can be broken down into the interactions among autonomous agents, i.e., AV-AV interactions, as well as between autonomous agents and human drivers, i.e., human-AI interactions. Decoupling the two enables us to systematically study the interactions between human drivers with ambiguous SVO and our autonomous agents. We refer to an autonomous agent's altruism toward a human as *sympathy* and define *cooperation* as the altruistic behavior among autonomous agents. Our rationale for decoupling the components of altruism is that they differ in nature. As an instance, sympathy may not be reciprocal as humans are heterogeneous in their SVO but cooperation among autonomous agents is essentially homogeneous, assuming that they hold the same SVO. We investigate each component of altruism separately to better understand the emerging behaviors and the mechanics of inducing altruism in autonomous agents. Following this definition, we can rewrite Eq. (7) as,

$$\begin{aligned} R_i &= r_i \cos\varphi_i + \left(\sin\vartheta_i R_i^{\mathrm{AV}} + \cos\vartheta_i R_i^{\mathrm{HV}}\right)\sin\varphi_i \\ &= \underbrace{r_i \cos\varphi_i}_{\text{egoistic term}} + \underbrace{\sin\vartheta_i \sin\varphi_i R_i^{\mathrm{AV}}}_{\text{cooperation term}} + \underbrace{\cos\vartheta_i \sin\varphi_i R_i^{\mathrm{HV}}}_{\text{sympathy term}} \end{aligned} \tag{12}$$

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 autonomous and human-driven vehicles, respectively, as perceived from the ith agent's perspective. We expand on this topic in Section V-C where we introduce the distributed reward structure.

C. Decentralized Reward Structure

Following the notions of sympathy and cooperation and the notation of Eq. (12) we decompose the decentralized reward received by agent $I_i \in I$ as,

$$\begin{aligned} R_i(s_i, a_i) &= R^{\mathrm{E}} + R^{\mathrm{C}} + R^{\mathrm{S}} \\ &= r_i(s_i, a_i)\cos\varphi_i \\ &\quad + \sin\vartheta_i \sin\varphi_i \sum_{j \in \mathrm{AV}} r_j(\tilde{\mathbf{o}}_i) + \sum_{j \in M} r_j(\tilde{\mathbf{o}}_i) \\ &\quad + \cos\vartheta_i \sin\varphi_i \sum_{h \in \mathrm{HV}} r_{i,h}^j(\tilde{\mathbf{o}}_i) + \sum_{j \in M} r_j(\tilde{\mathbf{o}}_i) \end{aligned} \tag{13}$$

in which $j \in I \setminus \{I_i\}$, $h \in (V \cup \{M\}) \setminus (I \cap \{M\})$. The r_i term denotes the ego vehicle's driving performance derived from metrics such as distance traveled, average speed, and a negative cost for changes in acceleration to promote a smooth and efficient movement by the vehicle. The cooperative reward term, $r_{i,j}^{\mathrm{AV}}$ accounts for the utility of the ego's allies. It is important to note that the ego vehicle only requires the observation $\tilde{\mathbf{o}}_i$ to compute R^{C} and not any explicit coordination or knowledge of the actions of the other agents. The sympathetic reward term, $r_{i,h}^{\mathrm{HV}}$ is defined as

$$r_{i,h}^{\mathrm{HV}} = \sum_h \frac{u_h}{\eta d_{i,h}^{\psi}} \tag{14}$$

where u_h denotes an HV's utility, e.g., its speed, $d_{i,h}$ is the distance between the observer autonomous agent and the hth HV, and η and ψ are dimensionless coefficients. The sympathetic reward term in Equation (14) acts as a proxy to put more importance on the state of the vehicles that are geographically closer to the ego-vehicle. ψ and η are hyper-parameters that allow us to control this proxy in order to achieve an optimal point at which the ego-agent both considers the behavior of the vehicles on the farther horizon and also puts more importance on its immediate neighbors. Moreover, the sparse scenario-specific *mission reward* term r_e^{M} in the case of our driving scenario is representing the success or failure of the merging maneuver,

$$r_e^{\mathrm{M}} = \begin{cases} 1/2, & \text{if } I_e \equiv M \text{ and merge is successful} \\ 0, & \text{o.w.} \end{cases} \tag{15}$$

D. Deep MARL for Sympathetic and Cooperative Driving

Two cascade multi-layer perceptron (MLP) networks are utilized as the feature extractor network (FEN) and the function approximator network (FAN), each with two layers of size 256 and 128 neurons, respectively, and rectified linear unit (ReLU) non-linearities. As introduced in Section V-A, the temporal information in a vehicle's observations is captured through integrating the history of the past actions in the observations and the feature extractor network must be able to efficiently extract meaningful patterns from this information. Both networks are trained end-to-end to enforce the feature extractor network to extract the most vital information that is required for estimating the state-action value function. The policy is trained offline and deployed to all agents to be executed in a distributed and online fashion, meaning that each agent makes independent decisions based on its observation but they all follow the same stochastic policy.

As we elaborated in Section III, the non-stationarity of the environment is a major problem in the concurrent training of multiple RL agents. We employ a semi-sequential training and policy dissemination algorithm to cope with this challenge and stabilize the training process. Algorithm 1 summarizes our overall methodology which is done in two stages. First, an experience replay buffer (ERB) is filled with data from simulation episodes, and then, random samples drawn from this buffer are used for updating the weights of both FEN and FAN networks. For simplicity, we refer to the set of all weights for both neural networks as $\mathbf{w}$. We use a novel method for scoring the entries in ERB and drawing them with a probability proportional to that score.

ERB is highly skewed due to the nature of our highway merging scenario. To elaborate, each episode can be morphologically broken down into two parts, straight driving on the highway and the merging point. The former mostly provides information and training samples that are useful for learning the basics of driving and the latter contains important information regarding inter-agent coordination and altruistic behavior, which is of our interest. Only a few time steps of each episode contain the merging point and the rest is mostly related to highway cruising. To balance the training data drawn from the experience replay, we randomly draw samples with a probability p_{ERB} proportional to their spatial distance from the merging point. This method showed better performance when compared to the most common method of prioritizing the experience replay based on a sample's last resulting reward.

After drawing a training sample from ERB, the agent $I_i \in I$ performs k_{diss} iterations of training while the weights $\mathbf{w}_j^-$ of all other agents $I_j (j = i)$ is frozen. The updated weights $\mathbf{w}_i^+$ are then disseminated to the other agents to update their policy. This process is then repeated for all agents until convergence. Doing so enables us to stabilize the training and train all agents concurrently. The key idea is to apply incremental updates and keep the environment stationary in between the updates so that the optimizer achieves convergence. This semi-sequential algorithm is illustrated in Figure 4 and Algorithm 1.

Different from other works, we borrow the notion of SVO from psychology to quantify the agent's degree of selfishness or altruism and, we frame our problem as a MARL problem that learns from interaction to compute an optimal altruistic policy; we proposed a data-driven framework that incorporates a well-engineered decentralized SVO reward structure to model cooperation and sympathy and use a suitable deep reinforcement learning architecture. Our proposed multi-agent training framework and policy dissemination process help to mitigate the non-stationarity problem in simultaneous multi-agent training while optimizing for a social utility. To summarize our design, as depicted in Figure 4 and explained in Algorithm 1, we train stochastic policies for altruistic agents in an online RL fashion. All the agents are trained concurrently in a simulation environment where they interact with each other and with the human-driven vehicles in the scene. Agents can share their policies in a semi-sequential manner only during the training phase and there is no coordination among agents after deployment, i.e., during inference. Agents are expected to take stand-alone decisions with no explicit information sharing among themselves and with HVs.

Algorithm 1 Semi-Sequential Multi-Agent Q-Learning

```
Initialize experience replay buffer (ERB) of size N_buff
for Episode = 1 to N_episode do
  Initialize episode with l_M(t_0) and v_M(t_0)
  for t = 1 to T_episode do
    Fill ERB with the tuples ([o_i], [a_i], [o_i'], [R_i])
    Calculate the relevance factor p_ERB for each entry in ERB
Initialize Q(s, a; w) with random weights w^-
Initialize target network ŵ with weights ŵ = w^-
for Frame = 1 to N_episode × T_episode do
  c_target = 0
  for I_i in I do
    Freeze the weights w^- for I_j where j = i
    for k = 1 to k_diss do
      Calculate the spatial distance
      Draw a sample from ERB based on p_ERB values
      w^+ ← w + α∇_w L(w)
      c_target++
      if c_target == n_target then
        ŵ ← w^+
    w^- = w^+ for all I_i ∈ I
```

VI. Implementation Details

We start this section with the 2D micro-traffic simulator we employed to generate simulation episodes and formulate the human driver model that imitates the behavior of an HV in mixed-autonomy environments. Practical details of training and validation are discussed before presenting our results in the next section.

A. Driving Simulator

We modified an OpenAI Gym environment [41] to enable multi-agent training and distributed execution in a mixed-autonomy highway merging scenario. The meta-actions

determined by the stochastic policy are translated to low-level steering and acceleration control signals through a closed-loop proportional–integral–derivative (PID) controller. The motion of the vehicles is then governed by a Kinematic Bicycle Model that determines the vehicles' yaw rate and acceleration. As a common practice in robotics, road segments and the motion of the agents are expressed in Frenet-Serret coordinates and broken into lateral and longitudinal movements.

In order to ensure learning generalizable policies rather than memorizing a sequence of actions by the function approximator network, the initial state of each simulation episode is randomized. This episode initialization is particularly critical as the resulting initial states must be still meaningful and valid for our desired conflictive highway merging scenario. Trivial episodes where the merging vehicle can easily merge into the highway regardless of the AVs' actions or the episodes where the AVs' do not have an opportunity to enable safe merging, not only do not add valuable information to the training process but also can lead to misleading measures. The initial longitude and speed of the cruising vehicles are uniformly randomized and the initial longitude $l_M(t_0)$ and speed $v_M(t_0)$ of the merging vehicle are drawn from a clipped-Gaussian distribution $N(x; \mu, \sigma, \delta)$ defined as,

$$N(x) = N(x; \mu, \sigma)\, \mathbb{1}(x - \mu + \delta) - \mathbb{1}(x - \mu - \delta) \quad (16)$$

where $N(x; \mu, \sigma)$ denotes a Gaussian distribution and $\mathbb{1}$ is the Heaviside step function. We elaborate on initializing episodes via parameters μ, σ, and δ in Section VII-E.

B. Human Driver Model

Lateral and longitudinal movements of HVs are mimicked by human driver models proposed by Treiber *et al.* and Kesting *et al.* [42], [43]. The lateral actions of HVs, i.e., the decision to perform a lane change, follow the Minimizing Overall Braking Induced by Lane changes (MOBIL) strategy [43]. MOBIL model allows a lane change only if the resulting acceleration $\text{acc}_n > -b_{\text{safe}}$ meets the safety criterion, and the incentive criterion is also satisfied,

$$\text{acc}_e^0 - \text{acc}_e + \sin\varphi_e(\text{acc}_n^0 - \text{acc}_n) + (\text{acc}_o^0 - \text{acc}_o) > \text{acc}_{\text{th}} \quad (17)$$

with acc_e, acc_n, and acc_o being the acceleration of the ego HV, the following vehicle in the target lane, and the following vehicle in the current lane, respectively, and acc_e^0, acc_n^0, and acc_o^0 are the corresponding accelerations assuming the ego HV has performed the lane change. acc_{th} is the threshold that determines if the ego HV shall perform the lane change. HV's SVO angle φ_e is also referred to as the politeness factor in the literature and is extracted from the empirical probability distribution illustrated in Figure 2.

The longitudinal acceleration of HVs follows the Intelligent Driver Model (IDM) [42]. The longitudinal Frenet acceleration of an HV, $\dot{l}_{\text{IDM}}$, is determined by

$$\dot{l}_{\text{IDM}} = \text{acc}_{\max}\left[1 - \left(\frac{\dot{l}}{v^{\text{set}}}\right)^4 - \left(\frac{d^{\text{des}}(l, \dot{l})}{d}\right)^2\right]. \quad (18)$$

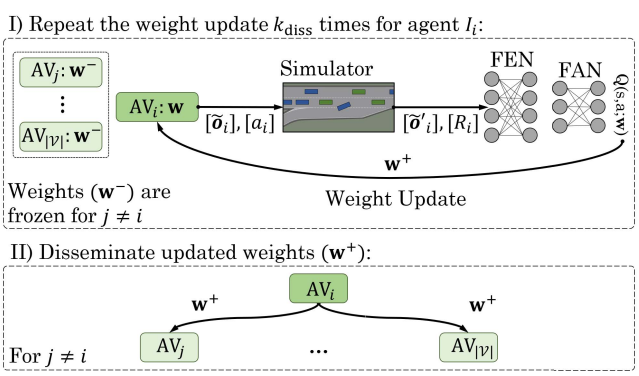


Fig. 4. Multi-agent training and policy dissemination process.

where $\dot{l}$ denotes the longitudinal Frenet speed of the HV, and the desired Frenet distance to the leading vehicle is controlled by d^{des}, defined as,

$$d^{\text{des}}(\dot{l}, \Delta\dot{l}) = d_0 + \dot{l}T_{\text{set}} + \frac{\dot{l}\Delta\dot{l}}{2\sqrt{\text{acc}_{\max}.\text{acc}_{\text{des}}}} \quad (19)$$

in which $\Delta\dot{l}$ is the approach rate, and the model parameters v_{set}, T_{set}, d_0, $\text{acc}_{\max}$, and acc_{des} are set speed, set time gap, minimum gap distance, maximum acceleration, and the desired acceleration, respectively. Additionally, the acceleration of the vehicle is a random variable defined as,

$$\dot{l} = \dot{l}_{\text{IDM}} + \frac{\sigma_{\text{vel}}}{\Delta t} N(0, 1) \quad (20)$$

with $N(0, 1)$ being a standard Gaussian random variable and σ_{vel} is the standard deviation of the velocity noise at the time step Δt of the simulation.

C. Training and Hyperparameter

The autonomous agents are trained using the semi-sequential multi-agent Q-learning algorithm that we introduced in Figure 4 and Algorithm 1 for 15,000 episodes that are generated by the procedure discussed in Section VI-A. The training process is repeated and compared across multiple runs to assure the stability of training and that it converges to similar policies every time. The trained policies are then evaluated for 2,000 randomized novel test episodes to gauge their efficacy. Test episodes are intentionally generated with a different and broader initialization range than the training episodes to demonstrate that agents actually are able to learn generalizable policies and not only memorize sequences of actions.

To guarantee that agents reach similar policies in terms of training stability, we conducted many rounds of training, and compute the average reward. To evaluate convergence, we use the average reward over multiple runs and take an average across runs. Figure 2 illustrates the training performance in terms of average reward and distance traveled. Despite the empirical success of the DQN and its variants, there is no guarantee for convergence. While we do not provide a convergence theoretical proof, the smoothness of the learning curve and our experiments shows that our method reaches

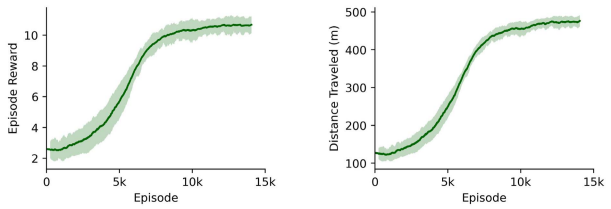


Fig. 5. Training performance of altruistic agents using the proposed algorithm.

TABLE I
COMPUTATION TIME ON DIFFERENT HARDWARE PLATFORMS

Computing platform	Computation time
NVIDIA Tesla V100 GPU	18 ms
OnLogic Karbon 700 x2	122 ms
NVIDIA GEFORCE GTX 1060	198 ms
NVIDIA Jetson AGX Xavier GPU	371 ms

stable policies and is able to train AVs that learned how to drive by using a decentralized reinforcement learning signal.

Using a GPU NVIDIA Tesla V100, our approach typically requires 2.7GB of memory for 4 agents and 20 HVs. While each run of 15,000 training episodes in the Tesla V100 GPU takes approximately 11h, a forward pass during testing for 4 AV takes 18ms. A total of 3,600 GPU were used for our experiments. Our simulation and training hyper-parameters are listed in Table IV We evaluate the time performance of our architecture in terms of computation time on a variety of hardware platforms. The results are presented in Table I. For instance, an online forward pass of the network utilizing an NVIDIA GEFORCE GTX 1060, will take around 198 ms.

VII. EXPERIMENTAL RESULTS

We break down the research questions of our interest into experimental hypotheses and investigate them through our experiments and ablation studies in this section.

A. Manipulated Variables

The two key variables in Eq. (13) are φ and ϑ that determine the level of altruism, which is the general term we use for both HVs and AVs, as well as the level of sympathy, which is the term for altruism toward HVs only. Our experiments are done in 2×6 settings with different values of *phi* and ϑ. Furthermore, we experiment with both autonomous, $M \in I$, and human-driven, $M \in V$, mission vehicle. Our experiment settings are:

- **HV+E.** autonomous agents are egoistic ($\varphi_i = 0$ for $I_i \in I$), and the mission vehicle is HV ($M \in V$);
- **HV+C.** autonomous vehicles are cooperative only ($\varphi_i = \varphi^{\boxdot}$ and $\vartheta_i = \pi/2$ for $I_i \in I$), and the mission vehicle is HV ($M \in V$);
- **HV+SC.** autonomous vehicles are sympathetic and cooperative ($\varphi_i = \varphi^{\boxdot}$ and $\vartheta_i = \pi/4$ for $I_i \in I$), and the merging vehicle is HV ($M \in V$);
- **AV+E/C/SC.** Duals of the above cases with autonomous mission vehicle ($M \in I$).

In **HV+SC** and **AV+SC** scenarios where autonomous agents have both sympathy and cooperation components, we set the sympathy angle to $\vartheta = \pi/4$ for the sake of fairness and to avoid imposing bias between HVs and AVs as they both carry humans or goods and neither should have a pre-assumed advantage over the other. The SVO angle φ is however tuned to reach the optimal level of altruism, we elaborate on this topic in Section VII-D and derive the optimal SVO angle $\varphi^{\boxdot}$.

B. Performance Measures

To gauge the impact of the aforementioned manipulated variables and other configurable parameters, 3 metrics are chosen that despite being correlated with each other, provide different insights on the efficacy of our solution. As a traffic-level metric, the average distance traveled by HVs and AVs is logged during simulation episodes. Additionally, counting the percentage of the episodes that experience a successful merge enables us to probe the overall social importance of a solution. Safety is also gauged by counting the percentage of episodes that contain at least one crash.

C. Hypotheses

The social and individual performance of altruistic and purely egoistic agents are compared through the 3 key hypotheses:

- **H1.** *While egoistic AVs fail to account for a merging HV, AVs that hold both sympathy and cooperation elements explore ways to enable safe and seamless merging. Therefore, we expect **HV+SC** to outperform **HV+E** and **HV+C** settings.*
- **H2.** *AVs with $\varphi = 0$ are able to implicitly learn the SVO of HVs and guide them to improve the overall performance of the group.*
- **H3.** *There exists a social value orientation angle $0 < \varphi^{\boxdot} < \pi/2$ for autonomous agents that can both lessen the number of crashes and improve the number of successful merges.*

D. Analysis and Results

1) Examining H1 : The main claim of hypothesis **H1** is the superiority of sympathetic cooperative AVs in creating socially optimal results when compared to egoistic autonomous AVs. To better understand the situation, we reiterate the driving scenario: the merging vehicle M, which can be either human-driven or autonomous, approaches a highway with a mixed group of HVs and AVs. M requires the cruising vehicles' assistance in order to be able to merge safely. Per our fundamental assumption, we do not rely on the HVs to compromise on their own utility as their SVO is unknown. Instead, it's on the AVs to create a safe corridor for M and, as we will show in Section VII-E, this goal cannot be achieved by a single AV alone and necessarily needs a cooperative action by the group of AVs.

Figure 6 illustrates an overall comparison between the settings defined in Section VII-A. Focusing on the cases with a human-driven merging vehicle, it is evident that in the

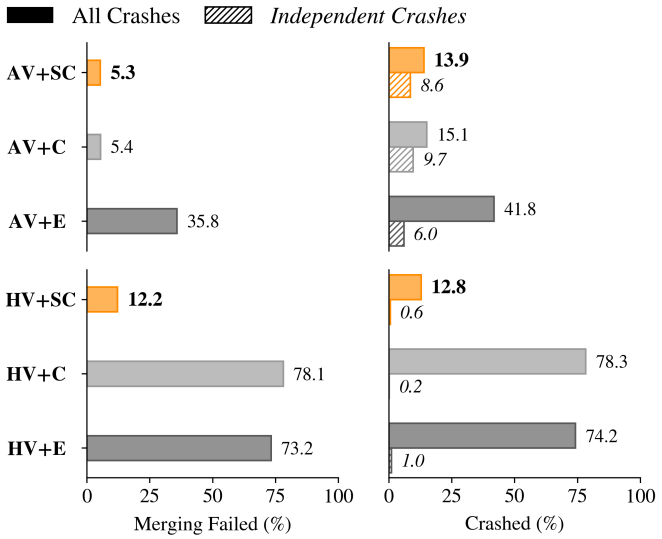


Fig. 6. Impact of *sympathy* and *cooperation* elements in traffic safety and success of the merging maneuver for both $M \in I$ or $M \in V$. Hatched bars show the number of independent crashes that do not involve the mission vehicle.

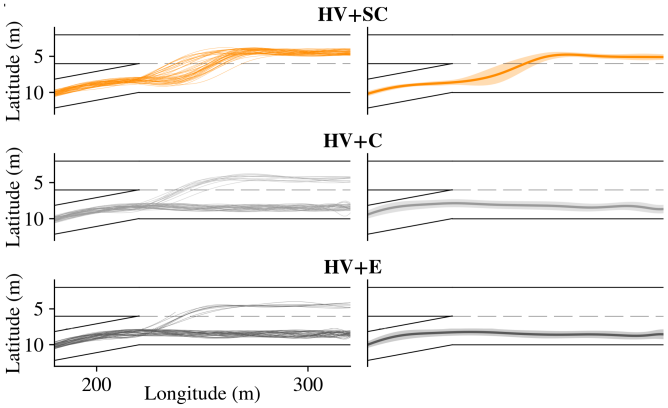


Fig. 7. Sampled trajectories of the human-driven mission vehicle M shows the efficacy of **SC** agents. Mean and standard deviation are shown on the right-hand side plots.

absence of the *sympathy* component in AVs, i.e., in **HV+E** and **HV+C** settings, merging fails in the majority of episodes. Failed merging leads to a crash in our simulator as vehicles cannot stop on the highway or the merging ramp and the merging vehicle that fails to merge collides with the barrier at the end of the merging ramp. This assumption is made to make our simulations more realistic and avoid unfeasible solutions that require full-stop on the highway. Therefore, most of the crash cases shown in Figure 6 are due to unsuccessful merging and not the lack of basic driving skills in HVs and AVs. As additional evidence, independent crashes that are not relevant to a failed merge are also plotted in Figure 6, which confirms the fact that the vehicles hold sufficient basic skills to maneuver on a highway and avoid collisions.

Figure 6 and 7 clarify the positive social impact that sympathy and cooperation make in terms of reducing the total number of crashes and failing to merge. However, a counter-argument against this comparison can be the fact that a rather conservative model is used to mimic HVs in our simulations and this might limit their capability in merging. To investigate this claim, we repeat the comparison with an autonomous mission vehicle that is more risk-tolerant and attempts more creative ways to merge into the highway. In the **AV+E** setting that AVs only care about their individual utility, although the results are better compared to **HV+E**, even the autonomous mission vehicle still fails to safely merge in more than 1/3 of the episodes. We conclude that our test case indeed creates a competitive and conflictive scene for the vehicles and showcases how incorporating sympathy and cooperation components in the reward structure of AVs leads to socially-desirable outcomes and improves safety and traffic flow. Figure 7 provides further intuition to this comparison by depicting a sampled set of mission vehicle's trajecto-ries in different experimental settings. It is evident that the un-sympathetic does not allow the mission vehicle to merge, causing its trajectory to end in the merging ramp.

2) Examining H2 : Figure 8 illustrates an example of autonomous agents trained with the sympathetic cooperative reward and a higher capacity neural network architecture. Although all AVs in this scenario work together to make the merging possible, we focus on the most impactful agent which is the "Guide AV" shown in orange color. Other AVs in this sample scenario (shown in green) compromise on their individual reward by accelerating, consuming more energy, and thus receiving less reward as defined in Section V-C. Interestingly, the Guide AV learns to first slow down and then change lane to left and open up space for M. After M successfully merges, the Guide AV finds its lane blocked by a HV so makes another lane change to the right and follows other AVs. Figure 8 demonstrates how AVs receive a significant reward when M merges into the highway. The reward structure defined in Section V-C contains multiple parameters, and the mission reward term r^M of Eq. (15) has a significant impact on the reward signal used to train our autonomous agents. In other words, the trained agents learn to take sequences of actions that lead to receiving r^M. This learning process includes learning to avoid collisions, navigating through the traffic, and if required affecting the behavior of other HVs.

As was emphasized before, autonomous agents do not have access to an explicit behavior model of human drivers and instead implicitly learn this model from experience during the training episodes. Although we employ a rather conservative model of human drivers to showcase our concept, it is expected that given sufficient training data, the autonomous agents can extract models of more complex human behaviors as well. However, the sensitivity of our solution to these models and the effect of human behaviors on inter-agent coordination is a topic worthy of investigation which we leave for our future work. As a relevant observation, AVs implicitly learn to predict the behavior of HVs and the fact that HVs commonly act egoistically (refer to Figure 2) and do not slow down for the merging vehicle. Hence, they do not rely on the HVs and instead compromise on their individual reward to enable the highway merging.

3) Examining H3: The experimental scenarios in Section VII-A are defined based on the optimal SVO angle φ^* of the autonomous agents. This parameter clearly has an important impact on the behavior of AVs and thus

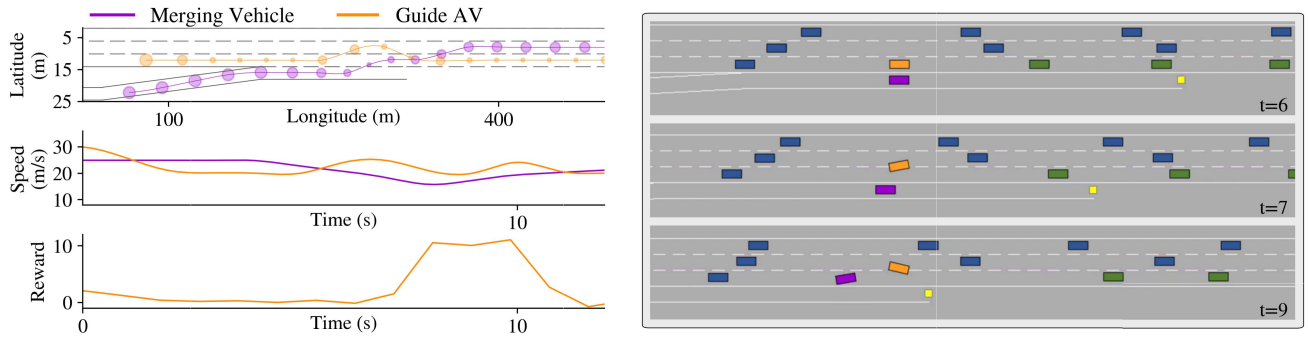


Fig. 8. An example of AVs (*green*) and HVs (*blue*) in **HV+SC** setting. Successful merging requires all the AVs to work together and none of them can achieve this goal alone. We focus on the "Guide AV" that makes the most significant impact and show how it learns to take sequences of actions to not only enable the mission vehicle to merge (by decelerating and performing a lane change to the left) but also manages to make the minimum compromise on its individual utility (by another lane change to right and cruising with optimal speed). The diameter of the circles on the trajectory plot shows the vehicles' speed.

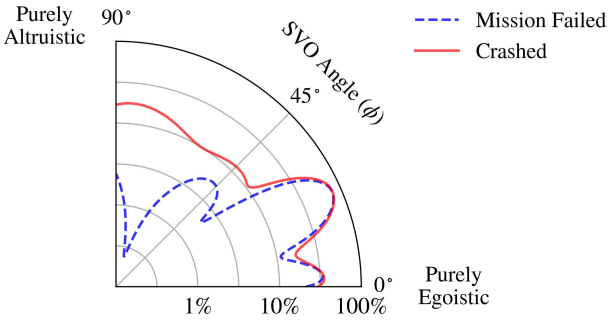


Fig. 9. Finding the optimal SVO angular phase $\varphi^{\boxed{?}}$ for AVs that results in the least number of crashes and failed merges. We performed several training runs and spent thousands of GPU hours sweeping a range of hyperparameters and SVO values and studying their impact on the output of driving scenarios. Using the optimization criteria in Eq. (21), we aim to choose an empirically optimal SVO value that optimizes both safety and efficiency of the driving scenarios.

TABLE II

NECESSITY OF MULTI-AGENT COORDINATION: A SINGLE SC AGENT IS NOT ABLE TO CREATE SOCIALLY-DESIRABLE OUTCOMES

	Mission Failed	Crashed	Distance Traveled
Single-agent (**HV+1SC**)	74.4%	74.5%	268.5m
Multi-agent (**HV+SC**)	**12.2%**	**12.8%**	**334.4**m

TABLE III

ABLATION STUDY ON REPRESENTING AGENT OBSERVATION $\tilde{\mathbf{o}}_i$

	Mission Failed	Crashed
Adding Autonomy Flag λ		
Without	5.0%	10.4%
With	**3.8%**	**8.9%**
Including Mission Vehicle o_M		
Without	4.2%	9.2%
With	**1.7%**	**8.3%**

the safety and traffic-flow metrics. We trained a large set of agents with different SVO angles and tested them in our case study driving scenario. The optimal SVO angle is then defined as the angle that results in the best performance metrics, i.e., the least number of episodes with collisions and failed merges. We formulate this simple optimization objective as the convex combination of the two metrics,

$$\varphi^{\boxed{?}} = \underset{\varphi}{\text{argmin}} \quad \xi . f_C(\varphi) + 1 - \xi . f_{MF}(\varphi) \tag{21}$$

where f_C and f_{MF} are the percentage of episodes with a crash and failed mission, respectively. The hyper-parameter ξ determines the importance of each performance metric and we choose it to be $\xi = 0.5$ as otherwise, it could bias the training process by putting more emphasis on either of the metrics. Figure 9 illustrates how the two metrics change when the autonomous agents' SVO is varied from $\varphi = 0$ (purely egoistic) towards $\varphi = \pi/2$ (purely altruistic). It is worth mentioning that neither of the two extremes seems optimal and a point between *caring about others* and *being selfish* leads to the most socially-desirable outcome.

A fair critique of the behavior of sympathetic cooperative agents can be the fact that the Guide AV, i.e., AV3 in Figure 3, decelerates and therefore slows down the group of vehicles behind only to allow the mission vehicle to merge. In other words, the utility of a big group of vehicles is being compromised for the sake of the mission vehicle. To investigate the fairness and effectiveness of this outcome, we measure the average distance traveled by HVs and AVs. Figure 10 reveals how despite the fact that in the **HV+SC** setting a group of vehicles needs to slow down to open up space for the mission vehicle, eventually both HVs and AVs manage to travel more distance when compared to a similar setup with egoistic agents (**HV+E**). It should be noted that the effect of Guide AV's deceleration gradually propagates through the platoon of vehicles behind and only affects a limited group of vehicles as the traffic in the platoon is not rigid and can contract and expand.

E. Ablation Studies

1) Necessity of Multi-Agent Coordination: Consider the highway merge scenario of Figure 3. Our claim is that all AVs require to work together to enable a safe and seamless merging and none of them can achieve this goal if the others do not cooperate. As elaborated in Section VI-A, we particularly

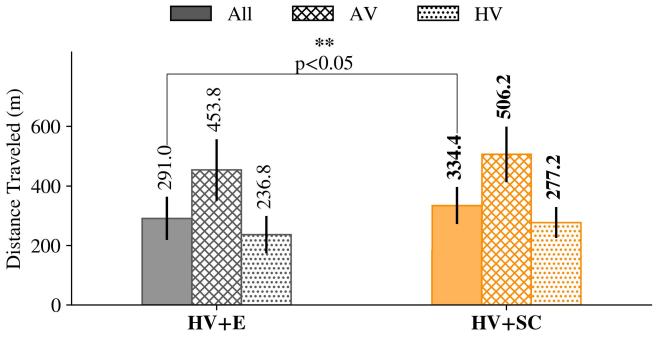


Fig. 10. Focusing on the highway merge problem, one valid question could be the fairness of slowing down the group of cruising vehicles on the Highway in order to enable assist the mission vehicle. Investigating the average distance traveled by vehicles (averaged over the length of the episode and all vehicles), reveals that despite making a compromise and slowing down a group of vehicles to allow the mission vehicle to merge, **SC** agents still lead to better overall traffic flow for both AVs and HVs.

design our scenarios to gauge the effectiveness of altruistic agents and inter-agent coordination. To complement our results in Figure 6 that back the hypothesis **H1**, we conducted an ablation study in the driving scenario of Figure 3 with the difference that only AV3 is sympathetic cooperative and label this scenario as **HV+1SC**. Table II demonstrates the necessity of multi-agent coordination and the fact that a single sympathetic cooperative AV, i.e., the Guide AV, is not able to achieve the mission of safe and seamless merging without help from the other AVs.

2) Designing Non-Trivial and Fair Scenarios : Our method for initializing simulation episodes is described in Eq. (16). Parameters μ and δ determine the range of the allowed values for the merging vehicle's initial longitude and speed. Trivial episodes that are too easy, i.e., always lead to successful merging, or too challenging, i.e., never result in a successful merge, can steer the training process in the wrong direction and must be avoided when initializing the episodes. Furthermore, the initial state of an episode can benefit different agents with various SVOs, and thus, one may argue that the superior performance of sympathetic cooperative agents as observed in Figures 6 and 7 is an artifact of the episode's initialization. We draw the initial values from a region that does not favor either of the social preferences. Two sets of parameters $(\mu_l, \delta_l, \sigma_l = 2\delta_l)$ and $(\mu_v, \delta_v, \sigma_v = 2\delta_v)$ are chosen for the initial longitude $l_M(t_0)$ and initial speed $v_M(t_0)$ of the merging vehicle, as listed in Table IV. Figure 11 illustrates the intuition behind choosing these values.

3) Observation-Space Representation : We discussed the details of how information is embedded into an agent's observation in Section V-A. Here we justify the design choices and show their positive impact on the performance. Table III shows the impact of including o_m in Eq. (9) as well as the autonomy flag λ of Eq. (10). Figure 12 summarizes the effect of integrating Π_j^A in Eq. (10), the history horizon h, and the type of the action encoding. We also experimented with sorting the rows of $\mathbf{o}^{(i)}$ in Eq. (9) based on vehicle ID and vehicles' longitude, as shown in Figure 12.

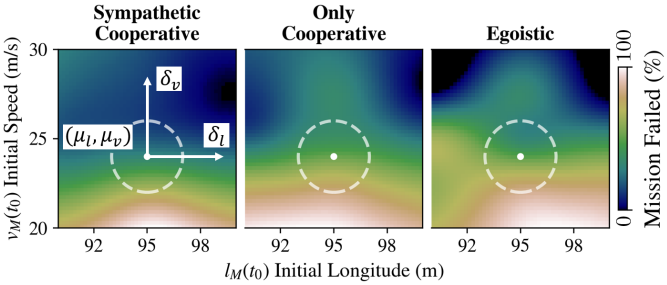


Fig. 11. Training episodes should not be trivial nor should they benefit a specific setting (**E**, **C**, **SC**). We ensure that we initialize our driving simulation episode in a region of parameters that are equally fair for all three settings. This has been done to ensure the sanity and validity of our experiments.

TABLE IV
TRAINING AND SIMULATION HYPER-PARAMETERS

Parameter	Value	Parameter	Value
N_{episode}	$15,000$	μ_l	$95m$
Batch size	32	δ_l	$2m$
N_{buffer}	$100,000$	μ_v	$24m/s$
α_0	0.0005	δ_v	$2m/s$
n_{target}	200	v_{set}	$25m/s$
Initial exploration ϵ_0	1.0	T_{set}	$0.5s$
Final exploration ϵ_f	0.1	d_0	$1m$
$\epsilon-$decay	Linear	$\text{acc}_{\max}$	$3m/s^2$
Optimizer	ADAM	acc_{des}	$-5m/s^2$
γ	0.95	acc_{th}	$0.2m/s^2$
$\lvert\mathcal{V}\rvert$	20	h	10
$\lvert\mathcal{I}\rvert$	4	ξ	0.5
T_{episode}	$18s$	k_{diss}	4

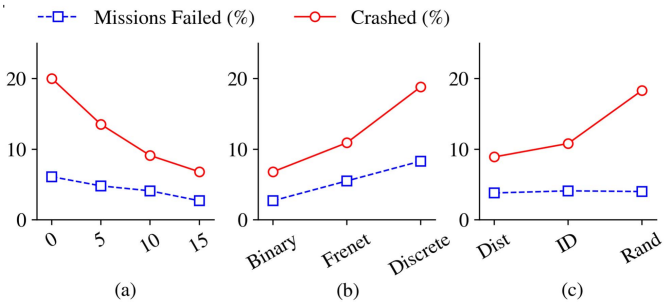


Fig. 12. *(a)* Length of action history h, *(b)* embedding type (Section V-A), *(c)* Sorting rows of Eq. (9) using longitudinal distance.

VIII. CONCLUDING REMARKS

A. Summary

Autonomous vehicles need to learn to co-exist with human-driven vehicles on the same road infrastructure. Deploying egoistic AVs that solely account for their individual interests on the road leads to sub-optimal and non-desirable social outcomes. In contrast, we compute the optimal SVO angle that optimizes the traffic metrics and demonstrates how altruistic AVs with the corresponding SVO can be trained to optimize a decentralized social utility that improves traffic flow, safety, and efficiency. We propose practical solutions to mitigate the non-stationarity problem in simultaneous multi-agent training and implicitly learn the behavior of human drivers from experience. Our experiments reveal that altruistic AVs are able to form alliances and affect the behavior of HVs in order to

create socially-desirable outcomes that benefit the group of the vehicles.

B. Limitations and Future Work

While this paper captures the fundamentals of social coordination and altruism in autonomous driving, many tangential aspects of the problem can be further studied. For example, we employed a conservative and limited model of human drivers. Although we expect our solution to be effective with other human behavior models as well, it is important to study its performance under different human behaviors. Also, the impact of communication imperfections and packet drops on inter-agent coordination can be further investigated using more complex communication models than those presented in this work. On the implementation side, more advanced neural architectures such as convolutional LSTM masked autoencoders and transformers can be leveraged to capture spatial and temporal information more effectively, a direction that we plan to explore in our future work.

REFERENCES

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[2] 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, 2016, pp. 1–9. [Online]. Available: http://www.roboticsproceedings.org/

[3] D. Sadigh, "Influencing interactions between human drivers and autonomous vehicles," in *Proc. U.S. Frontiers Eng. Symp.* Washington, DC, USA: National Academies Press, 2020, pp. 1–96.

[4] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus, "Social behavior for autonomous vehicles," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 50, pp. 24972–24978, 2019.

[5] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "Altruis-tic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2021, pp. 1–8.

[6] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "Cooperative autonomous vehicles that sympathize with human drivers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 4517–4524.

[7] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems," *Knowl. Eng. Rev.*, vol. 27, no. 1, pp. 1–31, Feb. 2012.

[8] N. Nardelli, G. Farquhar, T. Afouras, P. H. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.* 2017, pp. 1146–1155.

[9] A. Xie, D. Losey, R. Tolsma, C. Finn, and D. Sadigh, "Learning latent representations to influence multi-agent interaction," in *Proc. 4th Conf. Robot Learn. (CoRL)*, Nov. 2020, pp. 1–14.

[10] A. Shih, A. Sawhney, J. Kondic, S. Ermon, and D. Sadigh, "On the critical role of conventions in adaptive human-AI collaboration," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–14.

[11] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.

[12] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *Proc. Int. Conf. Auto. Agents Multiagent Syst.* Berlin, Germany: Springer, 2017, pp. 66–83.

[13] W. Z. Wang, M. Beliaev, E. Biyik, D. A. Lazar, R. Pedarsani, and D. Sadigh, "Emergent prosociality in multi-agent games through gift-ing," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1–9.

[14] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2681–2690.

[15] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *Proc. 17th Int. Conf. Mach. Learn.* Morgan Kaufmann, 2000, pp. 535–542.

[16] M. Kuderer, S. Gulati, and W. Burgard, "Learning driving styles for autonomous vehicles from demonstration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 2641–2646.

[17] H. N. Mahjoub, B. Toghi, and Y. P. Fallah, "A stochastic hybrid framework for driver behavior modeling based on hierarchical Dirichlet process," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–5.

[18] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer, "Imitating driver behavior with generative adversarial networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 204–211.

[19] E. Schmerling, K. Leung, W. Vollprecht, and M. Pavone, "Multimodal probabilistic model-based planning for human–robot interaction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3399–3406.

[20] B. Toghi *et al.*, "A maneuver-based urban driving dataset and model for cooperative vehicle applications," in *Proc. IEEE 3rd Connected Automated Vehicles Symp. (CAVS)*, Aug. 2020, pp. 1–6.

[21] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 1343–1350.

[22] L. Crosato, C. Wei, E. S. L. Ho, and H. P. H. Shum, "Human-centric autonomous driving in an AV-pedestrian interactive environment using SVO," in *Proc. IEEE 2nd Int. Conf. Hum.-Mach. Syst. (ICHMS)*, Sep. 2021, pp. 1–6.

[23] M. Van Vugt, R. M. Meertens, and P. A. Van Lange, "Car versus public transportation? The role of social value orientations in a real-life social Dilemma[1]," *J. Appl. Social Psychol.*, vol. 25, no. 3, pp. 258–278, 1995.

[24] J. Geary, H. Gouk, and S. Ramamoorthy, "Active altruism learning and information sufficiency for autonomous driving," 2021, *arXiv:2110.04580*.

[25] N. Tsoi, M. Hussein, J. Espinoza, X. Ruiz, and M. Vázquez, "SEAN: Social environment for autonomous navigation," in *Proc. 8th Int. Conf. Hum.-Agent Interact.*, Nov. 2020, pp. 281–283.

[26] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 797–803.

[27] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human–robot collaborative tasks," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2015, pp. 189–196.

[28] D. A. Lazar, E. Biyik, D. Sadigh, and R. Pedarsani, "Learning how to dynamically route autonomous vehicles on shared roads," 2019, *arXiv:1909.03664*.

[29] C. Wu, A. M. Bayen, and A. Mehta, "Stabilizing traffic with autonomous vehicles," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 6012–6018.

[30] C. Wu, A. Kreidieh, E. Vinitsky, and A. M. Bayen, "Emergent behav-iors in mixed-autonomy traffic," in *Proc. Conf. Robot Learn.*, 2017, pp. 398–407.

[31] E. Vinitsky *et al.*, "Benchmarks for reinforcement learning in mixed-autonomy traffic," in *Proc. Conf. Robot Learn.*, 2018, pp. 399–409.

[32] E. Biyik, D. Lazar, R. Pedarsani, and D. Sadigh, "Altruistic autonomy: Beating congestion on shared roads," 2018, *arXiv:1810.11978*.

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

[34] E. E. Marvasti, A. Raftari, A. E. Marvasti, Y. P. Fallah, R. Guo, and H. Lu, "Feature sharing and integration for cooperative cognition and perception with volumetric sensors," 2020, *arXiv:2011.08317*.

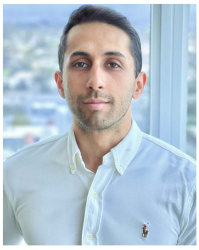
[35] R. Valiente, M. Zaman, S. Ozer, and Y. P. Fallah, "Controlling steering angle for cooperative self-driving vehicles utilizing CNN and LSTM-based deep networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 2423–2428.

[36] B. Toghi *et al.*, "Multiple access in cellular V2X: Performance analysis in highly congested vehicular networks," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Dec. 2018, pp. 1–8.

[37] W. B. Liebrand and C. G. McClintock, "The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation," *Eur. J. Personality*, vol. 2, no. 3, pp. 217–230, 1988.

[38] A. Garapin, L. Müller, and B. Rahali, "Does trust mean giving and not risking? Experimental evidence from the trust game," *Revue D'économie Politique*, vol. 125, no. 5, pp. 701–716, 2015.

[39] R. O. Murphy and K. A. Ackermann, "Social preferences, positive expectations, and trust based cooperation," *J. Math. Psychol.*, vol. 67, pp. 45–50, Aug. 2015.
[40] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," *Artif. Intell.*, vol. 299, Oct. 2021, Art. no. 103535.
[41] E. Leurent, Y. Blanco, D. Efimov, and O.-A. Maillard, "Approximate robust control of uncertain dynamical systems," 2019, *arXiv:1903.00220*.
[42] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E*, vol. 62, no. 2, p. 1805, 2000.
[43] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model MOBIL for car-following models," *Transp. Res. Rec.*, vol. 1999, pp. 86–94, Jan. 2007.



Behrad Toghi received the B.Sc. degree in electrical engineering from the Sharif University of Technology in 2016. He is currently pursuing the Ph.D. degree with the University of Central Florida. He has worked at the Honda Research Institute, Mercedes-Benz Research and Development North America and Ford Motor Company Research and Development from 2018 to 2022. His work is in the intersection of artificial intelligence and cooperative networked systems with a focus on autonomous driving.

Rodolfo Valiente received the B.Sc. degree from the Technological University Jose Antonio Echeverria in 2014 and the M.Sc. degree from the University of Sao Paulo (USP) in 2017. He is currently pursuing the Ph.D. degree in computer engineering with the University of Central Florida. His research interests include connected autonomous vehicles (CAVs), reinforcement learning, computer vision, and deep learning with a focus on the autonomous driving problem.

Dorsa Sadigh received the B.Sc. and doctoral degrees in electrical engineering and computer sciences from UC Berkeley in 2012 and 2017, respectively. She is an Assistant Professor with the CS and EE Departments, Stanford University. Her research interests include intersection of robotics, learning and control theory, and developing algorithms for safe and adaptive human–robot interaction.

Ramtin Pedarsani (Senior Member, IEEE) 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 the Ph.D. degree from the University of California, Berkeley, in 2015. He is an Assistant Professor with the ECE Department, University of California, Santa Barbara. His research interests include networks, game theory, machine learning, and transportation systems.

Yaser P. Fallah 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 at the Institute of Transportation Studies, University of California Berkeley, Berkeley, CA, USA. He is an Associate Professor with the ECE Department, University of Central Florida. His research sponsored by industry, USDoT, and NSF. His research interests include intelligent transportation systems and automated and networked vehicle safety systems.