2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

Stable and Efficient Shapley Value-Based Reward Reallocation for Multi-Agent Reinforcement Learning of Autonomous Vehicles

Songyang Han He Wang Sanbao Su Yuanyuan Shi Fei Miao

Abstract— **With the development of sensing and communication technologies in networked cyber-physical systems (CPSs), multi-agent reinforcement learning (MARL)-based methodologies are integrated into the control process of physical systems and demonstrate prominent performance in a wide array of CPS domains, such as connected autonomous vehicles (CAVs). However, it remains challenging to mathematically characterize the improvement of the performance of CAVs with communication and cooperation capability. When each individual autonomous vehicle is originally self-interest, we can not assume that all agents would cooperate naturally during the training process. In this work, we propose to reallocate the system's total reward efficiently to motivate stable cooperation among autonomous vehicles. We formally define and quantify how to reallocate the system's total reward to each agent under the proposed transferable utility game, such that communication-based cooperation among multi-agents increases the system's total reward. We prove that Shapley value-based reward reallocation of MARL locates in the core if the transferable utility game is a convex game. Hence, the cooperation is stable and efficient and the agents should stay in the coalition or the cooperating group. We then propose a cooperative policy learning algorithm with Shapley value reward reallocation. In experiments, compared with several literature algorithms, we show the improvement of the mean episode system reward of CAV systems using our proposed algorithm.**

I. INTRODUCTION

The rapid evolution of ubiquitous sensing, communication, and computation technologies has contributed to the revolution of cyber-physical systems (CPSs). Increasingly, multi-agent reinforcement learning (MARL)-based methodologies are integrated into the control process of physical systems and demonstrate prominent performance in a wide array of CPS domains. Connected and autonomous vehicles (CAVs) are one type of networked CPSs and multi-agent systems, with the development of vehicle-to-everything (V2X) communication technologies. The U.S. Department of Transportation (DOT) has estimated that DSRC (dedicated short-range communications) based V2V communication can address up to 82% of all crashes in the U.S. and potentially save thousands of lives and billions of dollars [1]. Information sharing of basic safety messages (BSMs) benefits CAVs coordination and control approaches in scenarios such as cross intersections or lane-merging [2], [3], [4], platoon and adaptive cruise control (ACC) [5], [6], [7].

However, it remains challenging to formally characterize the improvement of learning-based decision-making for CAV systems with V2X connections and to reallocate the system's total reward efficiently to motivate stable cooperation of individual autonomous vehicles. Existing learning-based planning or control approaches do not utilize communication or potential shared information for autonomous vehicles yet [8], [9], [10], [11]. How V2X communication benefits MARL, improves the system's total reward, and motivates cooperation has not been formally defined or quantified.

We propose to formally define and quantify the value of communication-based cooperation to MARL based on Shapley value [12], [13], and use it to reallocate the system's total reward to the individual agent to motivate the cooperation. It will build a research foundation to formally quantify the value of information sharing and motivate cooperation for the CAV research society, and the results can be leveraged to other networked CPSs to better understand the benefits of cooperation. Our proposed approach includes the following three major novelties and contributions.

- We formally define and quantify how to efficiently reallocate the system's total reward to the individual agent to motivate information sharing and stable cooperation among multi-agents.
- We define a transferable utility game formulation to study the reward reallocation problem, and prove that Shapley value-based reward reallocation is efficient in Theorem 1. We prove that the reward reallocation scheme is stable if the transferable utility game is a convex game in Theorem 2. Hence, each individual agent should cooperate to receive more rewards.
- We propose a cooperative policy learning with Shapley value reward allocation algorithm. In experiments, we show the improvement of the system's total reward, velocity, and comfort of CAV systems using the Shapley value of the grand coalition (all agents cooperate).

II. RELATED WORK

Cooperative games attract increasing research interests in MARL [14]. Existing works have been investigated to encourage every agent to work collaboratively by assigning rewards appropriately, such as a value decomposition network [15], [16], subtracting a counterfactual baseline [17], or an implicit method [18]. Multi-agent deep deterministic policy gradient (MADDPG) applies a centralized Q-function

This work was supported by NSF 1849246, NSF 1952096, and NSF 2047354 grants. S. Han, S. Su, F. Miao are with the Department of Computer Science and Engineering, University of Connecticut, Storrs Mansfield, CT, USA. Email: {songyang.han, sanbao.su, fei.miao}@uconn.edu. H. Wang is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. Email: {wanghe}@shanghaitech.edu.cn. Y. Shi is with the Electrical and Computer Engineering Department, University of California, San Diego, La Jolla, CA, USA. Email: {yyshi}@eng.ucsd.edu.

to address the problem caused by the non-stationary environment [19]. Its scalability can be improved by adding an attention mechanism to the centralized critic [20]. Distributed execution with communication among agents with centralized critics methods and online learning with communicative actions [21], [22], [23], [24], [25], [26] have shown performance improvement over non-communicative agents. In all these works, agents are assumed to cooperate during the training process, however, whether agents will cooperate or whether the cooperative coalition is **stable** has not been answered yet.

Another related line of work is coalition formation games, where selfish and rational agents choose to participate in coalitions to maximize their own utility. There are two classes of coalition formation games: static and dynamic formation processes. In the former class, several works employ bargaining methods [27], [28] to achieve the agreement and no agents would deviate from the agreement once it begins. Besides, some works rely on concepts of cooperative game theory to acquire stable coalitions and fair payoff distribution methods, such as Shapley value [29], core [30], [29] and kernel [31]. However, these methods do not apply to our setting because we consider a dynamic process. Methods that be used for analyzing dynamic coalition formation include Bayesian RL and MARL [32], [33]. But these works all use MARL as a tool to analyze the coalition formation game, while we consider the converse - we leverage analytical tools in cooperative game and coalition formation game for reward reallocation design in MARL.

Shapley value [12] is one classical method to divide the total payoff of all players in a cooperative game so that each player receives his or her fair payment. It uniquely provides an equitable assignment of values to agents and is also one metric to measure the importance and marginal contribution of each player to the system. Recent literature on learning has developed a principled framework of "data Shapley" [34] to address data valuation and "feature Shapley" [35] to measure the importance of features to supervised learning algorithms. Computing Shapley is still challenging and requires computing all the possible marginal contributions, which is exponentially large in the training process. Therefore, approximation methods to estimate the Shapley value of agents, data, and features have been designed and are presented in the game theory and machine learning literature [13], [36], [34], [35]. However, in this work, we leverage the exploration and exploitation power of reinforcement learning to reduce the computational complexity significantly by avoiding traversing all possible marginal contributions.

III. Preliminary

A. Convex Game

In cooperative games, a Transferable Utility game (**TU game**) with n agents can be represented by a pair $(\mathcal{N}, v)$, where $\mathcal{N} = \{1, \ldots, n\}$ is a set of agents and $v : 2^n \to \mathbb{R}$ is the real-valued characteristic function [37]. More specifically, $\mathcal{N}$ is the grand coalition in which all the n agents cooperate together. For any coalition $\mathcal{C} \subseteq \mathcal{N}$, $v(\mathcal{C})$ measures its value. A TU game (N, v) is a **convex game** if for any $\mathcal{C}, \mathcal{D} \subseteq \mathcal{N}$, $v(\mathcal{C} \cup \mathcal{D}) \geq v(\mathcal{C}) + v(\mathcal{D}) - v(\mathcal{C} \cap \mathcal{D})$. An **outcome** of any TU game consists of two parts, a **coalition structure** $\mathcal{CS}$ and a **payoff vector** $x \in \mathbb{R}^n$, denoted as a pair $(\mathcal{CS}, x)$. The coalition structure $\mathcal{CS} = \{\mathcal{C}^1, \ldots, \mathcal{C}^k\}$ over $\mathcal{N}$ is a possible partition of $\mathcal{N}$ and the corresponding payoff vector $x = (x^1, \ldots, x^n) \in \mathbb{R}^n$ satisfies: (1) $x^i \geq 0$ for each $i \in \mathcal{N}$ and (2) $x(\mathcal{C}^j) \leq v(\mathcal{C}^j)$, where $x(\mathcal{C}^j) = \sum_{i \in \mathcal{C}^j} x^i$, for any $j \in \{1, \ldots, k\}$.

B. Solution Concepts

There are several solution concepts that identify sets of outcomes with appealing properties, such as stability and efficiency. In this paper, we investigate four widely used solution concepts [37], as follows.

Definition 1 (**Core and Stable Outcome**). *Given a TU game $(\mathcal{N}, v)$, the Core is the set of all stable outcomes $(\mathcal{CS}, x)$ such $x(\mathcal{C}) \geq v(\mathcal{C})$ for every $\mathcal{C} \subseteq \mathcal{N}$, where $x(\mathcal{C}) = \sum_{i \in \mathcal{C}} x^i$.*

Definition 2 (**Efficient Outcome**). *If $x(\mathcal{C}) = v(\mathcal{C})$ for any $\mathcal{C} \in \mathcal{CS}$ in an outcome $(\mathcal{CS}, x)$, then this outcome is efficient and maximizes the social welfare.*

Definition 3 (**Shapley Value**). *Given a TU game $(\mathcal{N}, v)$, the Shapley Value for each agent $i \in \mathcal{N}$ is denoted by $\phi^i(v)$ and is given by*

$$\phi^i(v) = \sum_{\mathcal{C} \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathcal{C}|!(|\mathcal{N}| - |\mathcal{C}| - 1)!}{|\mathcal{N}|!} \left(v(\mathcal{C} \cup \{i\}) - v(\mathcal{C})\right).$$

C. Multi-agent Actor-Critic

The environment's state transition of MARL is influenced by the policy of all agents and it is non-stationary from a single agent's view. To alleviate this problem and stabilize training, MADDPG is proposed using a centralized Q-function that has global state information [19]. The ith agent maximizes its own expected return $J(\theta^i)$ and its gradient is

$$\begin{aligned} &\nabla_{\theta^i} J(\theta^i) = \\ &\mathbb{E}_{s,a \sim \mathcal{D}} \left[\nabla_{\theta^i} \pi^i(s^i) \nabla_{a^i} Q^i(s, a^1, ..., a^n)|_{a^i = \pi^i(s^i)} \right], \end{aligned} \tag{1}$$

where $Q^i(s, a^1, ..., a^n)$ is a centralized action-value function, s is the joint state, π^i is the policy. The experience replay buffer $\mathcal{D}$ contains $(s, a^1, ..., a^n, s', r^1, ..., r^n)$. The centralized critic Q^i is trained using the Bellman loss:

$$\begin{aligned} &\mathcal{L}(\theta^i) = \mathbb{E}_{s,a,r,s' \sim \mathcal{D}}[y - Q^i(s, a^1, ..., a^n)]^2, \\ &y = r^i + \gamma Q^{i\prime}(s', a^{1\prime}, ..., a^{n\prime})|_{a^{j\prime} = \pi^j(s^{j\prime})}, \end{aligned} \tag{2}$$

where $Q^{i\prime}$ is the target network and γ is a discount factor. Note that this algorithm adopts the centralized training and decentralized execution paradigm. When testing, each agent can only access its local state to select actions.

IV. Problem Formulation

We first formulate the connected autonomous vehicles (CAVs) problem as a MARL problem. We consider n agents (e.g., autonomous vehicles) in the agent set $\mathcal{N} = \{1, ..., n\}$. Each agent i is associated with an action $a^i \in \mathcal{A}^i$ and a

state $s^i \in \mathcal{S}^i$. The global joint state is $s = (s^1, ..., s^n) \in \mathcal{S} := \mathcal{S}^1 \times \cdots \times \mathcal{S}^n$. Similarly, the global joint action is $a = (a^1, ..., a^n) \in \mathcal{A} := \mathcal{A}^1 \times \cdots \times \mathcal{A}^n$. Each coalition has a stage-wise reward function $R(s, a, \mathcal{C})$, where we denote a subset of the vehicles $\mathcal{N}$ as a coalition $\mathcal{C}$. Specially, the $\mathcal{N}$ denotes the grand coalition in which all the n agents cooperate together. Let $R(s, a, \mathcal{N})$ to represent the global stage-wise reward under the grand coalition. For the empty coalition $\emptyset$, $R(s, a, \emptyset) := 0$. In our work, we assume that all rewards are nonnegative.

We consider each agent is associated with a localized policy $\pi^i(a^i|s^{\mathcal{C}})$ to choose a local action a^i given its local state $s^{\mathcal{C}}$. We use $\pi(a|s)$ to denote the joint policy. The system's objective is to find a policy $\pi(a|s)$ to maximize the system's discounted total reward under the grand coalition,

$$\max_{\pi(a|s)} \mathbb{E}_{a_t \sim \pi(a|s)} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}(s_t, a_t, \mathcal{N}) | s_0 = s, a_0 = a \right],$$

where γ is the discount factor. However, how to motivate agents to form a grand coalition has not been studied yet.

The MARL problem is usually solved by the centralized training and decentralized execution paradigm that is first proposed in [19]. During training, there is a centralized critic $Q(s, a)$ that has access to the global state s and global action a. This critic is used to train a localized actor for decentralized execution. However, having such a powerful centralized critic for autonomous vehicles is not easy, especially when each vehicle is self-interested and focuses to maximize its own reward (e.g., maximize its own driving speed or minimize its own travel time). Vehicles are not assumed to be fully cooperative. Therefore, the goal of this paper is to design a better total reward reallocation mechanism, under which agents are willing to collaborate with each other by sharing their state and action information.

Simply using the original action value function $Q(s, a)$ is not guaranteed to learn a policy where agents are willing to cooperate. In order to find a better total reward for policy learning, we formulate and analyze the game between CAVs using the cooperative game framework. We consider a Transferable Utility game (**TU game**) $G = (\mathcal{N}, v)$, where v is the characteristic function introduced in Section III-A. Consider the entire MARL problem across all timesteps, the value of any coalition $\mathcal{C}$ given the current state and action is

$$v(\mathcal{C}|s, a) = \mathbb{E}_{\pi(a|s)} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}(s_t, a_t, \mathcal{C}) | s_0 = s, a_0 = a \right].$$

Specially, when $\mathcal{C} = \mathcal{N}$ we have

$$v(\mathcal{N}|s, a) = \mathbb{E}_{\pi(a|s)} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}(s_t, a_t, \mathcal{N}) | s_0 = s, a_0 = a \right].$$

It shows the value of the grand coalition $\mathcal{N}$ is exactly the system's total reward. Note that the characteristic function is not the state value function V(s) used in reinforcement learning. The characteristic function is also a function of the coalition $\mathcal{C}$ in the TU game.

We want to find a stable and efficient total reward reallocation for this TU game such that agents can learn a cooperative policy. In the following section, we show how to utilize the Shapley value to solve this problem.

V. Stable and Efficient Reward Reallocation

We first define the Shaley value for the Transferable Utility game (TU game) $G = (\mathcal{N}, v)$ defined in Section IV and show that the Shaley value satisfies the axioms for a fair reward reallocation. We then present the Shapley value is an efficient reward reallocation in Theorem 1. Moreover, we show the Shapley value is a stable reward reallocation if the TU game $G = (\mathcal{N}, v)$ is a convex game in Theorem 2. Then we give an example of the convex game and design Algorithm 1. An efficient solution means the system's total reward is completely distributed to each agent for cooperative policy learning. A stable solution means any agent cannot get more payoff if they leave the coalition. For CAVs, a stable solution means that vehicles will stay within the coalition, communicate and cooperate with other coalition members to optimize the coalition-level objective.

We define the Shaley value of the TU game $G = (\mathcal{N}, v)$ as follows.

Definition 4 (**Shapley Value**). *The Shapley value of the TU game $G = (\mathcal{N}, v)$ is defined to be*

$$\phi^i(s, a) := \sum_{\mathcal{C} \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathcal{C}|!(n - |\mathcal{C}| - 1)!}{n!} [v(\mathcal{C} \cup \{i\}|s, a) - v(\mathcal{C}|s, a)]. \tag{3}$$

The Shaley value of the TU game $G = (\mathcal{N}, v)$ is extended for the multi-agent sequential decision problem from Definition 3. The intuition is that each agent's total reward is proportional to its total contribution to the entire system. It is used to quantify the contribution of each agent for the communication-based cooperation in MARL. Based on the definition of $\phi^i(s, a)$ in (3), it is straightforward to see that the Shaley value satisfies the axioms for a fair reward reallocation as follows:

Proposition 1. *The Shapley value of the TU game $G = (\mathcal{N}, v)$ satisfies the following axiomatic characterization for a fair reward reallocation:*

1) *Symmetric: if $v(\mathcal{C} \cup \{i\}|s, a) = v(\mathcal{C} \cup \{j\}|s, a)$ for any coalition $\mathcal{C} \subseteq \mathcal{N} \setminus \{i, j\}$, then $\phi^i(s, a) = \phi^j(s, a)$.*
2) *Dummy player: if $v(\mathcal{C} \cup \{i\}|s, a) = v(\mathcal{C}|s, a)$ for any coalition $\mathcal{C} \subseteq \mathcal{N}$, then $\phi^i(s, a) = 0$.*
3) *Additivity: for any two v^1 and v^2, $\phi^i(s, a|v^1 + v^2) = \phi^i(s, a|v^1) + \phi^i(s, a|v^2)$ for each i, where $\phi^i(s, a|v^1 + v^2)$ is Shapley value of the TU game $(\mathcal{N}, v^1 + v^2)$ and $(v^1 + v^2)(\mathcal{C}|s, a) = v^1(\mathcal{C}|s, a) + v^2(\mathcal{C}|s, a)$.*

If a reward reallocation satisfies the axioms of symmetric, dummy player, and additivity, it is called a fair reward reallocation [37]. Besides fairness, we show that the **outcome** $(\{\mathcal{N}\}, \phi)$ with Shaley value is an efficient outcome of the TU game $G = (\mathcal{N}, v)$, where $\{\mathcal{N}\}$ is the coalition structure for the grand coalition, and $\phi(s, a) = (\phi^1(s, a), ..., \phi^n(s, a))$ is

the payoff vector for each agent given the current state s and action a. The payoff $\phi^i(s,a)$ for each agent is the reallocated total reward. It is regarded as a new action value function after the reward reallocation.

Theorem 1. *The outcome $(\{\mathcal{N}\},\phi)$ with Shaley value is an efficient outcome of the TU game $G=(\mathcal{N},v)$ and the Shapley value is an efficient reward reallocation.*

Proof. Note that $v(\emptyset|s,a)=0$, we have

$$\begin{aligned}\sum_{i\in\mathcal{N}}\phi^i(s,a) &= \sum_{i\in\mathcal{N}}\sum_{\mathcal{C}\subseteq\mathcal{N}\setminus\{i\}}\frac{|\mathcal{C}|!(n-|\mathcal{C}|-1)!}{n!}[v(\mathcal{C}\cup\{i\}|s,a)\\ -v(\mathcal{C}|s,a)] &= n\frac{(n-1)!1!}{n!}v(\mathcal{N}|s,a)+n\frac{0!(n-1)!}{n!}v(\emptyset|s,a)\\ &+\sum_{\mathcal{C}\subset\mathcal{N},\mathcal{C}\neq\emptyset,|\mathcal{C}|=p}(p\frac{(p-1)!(n-p)!}{n!}-\\ &(n-p)\frac{p!(n-p-1)!}{n!})v(\mathcal{C}|s,a)\\ &= v(\mathcal{N}|s,a). \end{aligned}\tag{4}$$

Note that the first equation is a telescoping sum where both $v(\mathcal{N}|s,a)$ and $v(\emptyset|s,a)$ appear once for each agent. The value of coalition $\mathcal{C}$ with p agents appears p times with positive sign, once for each agent in $\mathcal{C}$; it also appears $n-p$ times with negative sign, once for each agent not in $\mathcal{C}$. □

Moreover, if the TU game $G=(\mathcal{N},v)$ is a convex game, we show that the Shaley value is a stable reward reallocation. A convex game is defined as follows:

Definition 5 (Convex Game). *The TU game $G=(\mathcal{N},v)$ is a convex game if for any pair of coalitions $\mathcal{C},\mathcal{D}\subseteq\mathcal{N}$,*

$$v(\mathcal{C}\cup\mathcal{D}|s,a)+v(\mathcal{C}\cap\mathcal{D}|s,a)\geq v(\mathcal{C}|s,a)+v(\mathcal{D}|s,a).$$

For a convex game, we have the following theorem that guarantees the core is nonempty, and there exists a stable solution for reallocating $v(\mathcal{N}|s,a)$ to each individual agent to motivate cooperation.

Theorem 2. *If the TU game $G=(\mathcal{N},v)$ is a convex game, the outcome $(\{\mathcal{N}\},\phi)$ with Shaley value is in the core and the Shapley value is a stable reward reallocation.*

Proof. Note that the Shapley value for the TU game $G=(\mathcal{N},v)$ is

$$\phi^i(s,a)=\sum_{\mathcal{C}\subseteq\mathcal{N}\setminus\{i\}}\frac{|\mathcal{C}|!(n-|\mathcal{C}|-1)!}{n!}[v(\mathcal{C}\cup\{i\}|s,a)-v(\mathcal{C}|s,a)].$$

We then show that Shapley value allocation lies in the core of the TU game $G=(\mathcal{N},v)$.

Let the payoff vector defined by the marginal contribution of each agent, i.e., $x^i=v(\mathcal{C}\cup\{i\}|s,a)-v(\mathcal{C}|s,a)$ for $i\in\mathcal{N}$. The goal is to prove that $x=(x^1,\ldots,x^n)$ is in the core. For any coalition $\mathcal{C}=\{i^1,i^2,\ldots,i^k\}\subseteq\mathcal{N}=\{1,...,n\}$, we can write $v(\mathcal{C})$ as follows by telescoping sum,

$$\begin{aligned}&v(\mathcal{C}|s,a)=v(\{i^1\}|s,a)-v(\emptyset|s,a)+v(\{i^1,i^2\}|s,a)\\ &-v(\{i^1\}|s,a)+\cdots+v(\mathcal{C}|s,a)-v(\mathcal{C}\setminus\{i^k\}|s,a).\end{aligned}\tag{5}$$

Without loss generality, we assume that $i^1<i^2<\cdots<i^k$. Then for each $j\in\{1,\ldots,k\}$, we let $\mathcal{D}=\{1,2,\ldots,i^{j-1}\}$ and by the definition of convex games,

$$\begin{aligned}&v(\{i^1,\ldots,i^{j-1},i^j\}|s,a)-v(\{i^1,\ldots,i^{j-1}\}|s,a)\\ &\leq v(\{1,\ldots,i^{j-1},i^j\}|s,a)-v(\{1,\ldots,i^{j-1}\}|s,a)\\ &=x^{i^j},\end{aligned}\tag{6}$$

where the last equation is based on $x^{i^j}=v(\mathcal{D}\cup\{i^j\}|s,a)-v(\mathcal{D}|s,a)$. By summing up (6) from $j=1$ to k, we have $v(\mathcal{C}|s,a)\leq\sum_{j=1}^k x^{i^j}=x(\mathcal{C}|s,a)$. Therefore, x is a stable solution in the core based on Definition 1 and any coalition $\mathcal{C}$ will not have incentives to deviate from the grand coalition. Moreover, since Shapley value is a convex combination of our constructed x, and core is a convex set, the outcome $(\{\mathcal{N}\},\phi)$ with Shaley value is also stable and lies in the core.

□

According to Theorem 1 and Theorem 2, the outcome $(\{\mathcal{N}\},\phi)$ with Shaley value is a stable and efficient outcome of the TU game $G=(\mathcal{N},v)$ if G is a convex game. No agent wants to deviate from the grand coalition in this case. Then, using the action value function $Q(s,a)$ for the policy learning is not the best choice. We can reallocate the system's total reward to each agent using the Shaley value $\phi^i(s,a)$. It is a fair, stable, and efficient total reward reallocation for a convex game. Now, we give an example for the convex game.

Proposition 2. *When $R(s,a,\mathcal{C})=\sum_{i\in\mathcal{C}}R(s,a,\{i\})$ for any nonempty coalition $\mathcal{C}\subseteq\mathcal{N}$, the TU game $G=(\mathcal{N},v)$ is a convex game.*

Proof. For any nonempty coalition $\mathcal{C}\subseteq\mathcal{N}$, $s\in\mathcal{S}$, and $a\in\mathcal{A}$, we have

$$\begin{aligned}v(\mathcal{C}|s,a)&=\mathbb{E}_{a_t\sim\pi(a|s)}\left[\sum_{t=0}^{\infty}\gamma^t R_{t+1}(s_t,a_t,\mathcal{C})|s_0=s,a_0=a\right]\\ &=\sum_{i\in\mathcal{C}}\mathbb{E}_{a_t\sim\pi(a|s)}\left[\sum_{t=0}^{\infty}\gamma^t R_{t+1}(s_t,a_t,\{i\})|s_0=s,a_0=a\right]\\ &=\sum_{i\in\mathcal{C}}v(\{i\}|s,a).\end{aligned}\tag{7}$$

Then we have for any pair of nonempty coalitions $\mathcal{C},\mathcal{D}\subseteq\mathcal{N}$, $s\in\mathcal{S}$, and $a\in\mathcal{A}$

$$\begin{aligned}&v(\mathcal{C}\cup\mathcal{D}|s,a)+v(\mathcal{C}\cap\mathcal{D}|s,a)\\ &=\sum_{i\in\mathcal{C}\cup\mathcal{D}}v(\{i\}|s,a)+\sum_{i\in\mathcal{C}\cap\mathcal{D}}v(\{i\}|s,a)\\ &=\left(\sum_{i\in\mathcal{C}}v(\{i\}|s,a)+\sum_{i\in\mathcal{D}\setminus\mathcal{C}}v(\{i\}|s,a)\right)+\sum_{i\in\mathcal{C}\cap\mathcal{D}}v(\{i\}|s,a)\\ &=\sum_{i\in\mathcal{C}}v(\{i\}|s,a)+\sum_{i\in\mathcal{D}}v(\{i\}|s,a)\\ &\geq v(\mathcal{C}|s,a)+v(\mathcal{D}|s,a).\end{aligned}\tag{8}$$

If the pair of coalitions $\mathcal{C}, \mathcal{D} \subseteq \mathcal{N}$ has at least one empty set, it is easy to check $v(\mathcal{C} \cup \mathcal{D}|s,a) + v(\mathcal{C} \cap \mathcal{D}|s,a) \geq v(\mathcal{C}|s,a) + v(\mathcal{D}|s,a)$ considering $v(\emptyset|s,a) = 0$. □

Proposition 2 gives us one example of the convex game. There exits other kinds of convex games, for example, when $R(s,a,\mathcal{C}) = -\max_{i \in \mathcal{C}} R(s,a,\{i\})$ [38] for any $\mathcal{C} \neq \emptyset$.

Algorithm 1: Cooperative Policy Learning with Shapley Value Reward Reallocation

1 Randomly initialize the characteristic network v and the actor network π^i for agent i. Initialize target networks $v', \pi^{i\prime}$;
2 **for** *each episode* **do**
3 Initialize a random process $\mathcal{X}$ for action exploration;
4 Receive initial state $\mathbf{s}$;
5 **for** *each timestep* **do**
6 Randomly sample a coalition $\mathcal{C}$. All agents in coalition $\mathcal{C}$ cooperate with each other at this time step.
7 For each agent i, select action $a^i = \pi^i(s^{\mathcal{C}}) + \mathcal{X}$ w.r.t the current policy and exploration. Execute actions $\mathbf{a} = (a^1, ..., a^n)$ and observe the reward $\mathbf{r}$ and the new state information $\mathbf{s}'$. Store $(\mathbf{s}, \mathbf{a}, \mathcal{C}, \mathbf{r}, \mathbf{s}')$ in replay buffer $\mathcal{D}$. Set $\mathbf{s} \leftarrow \mathbf{s}'$;
8 **for** *each agent* **do**
9 Sample a random minibatch of samples $(\mathbf{s}_k, \mathbf{a}_k, \mathcal{C}_k, \mathbf{r}_k, \mathbf{s}'_k)$ from $\mathcal{D}$;
10 Set $y_k = r_k^{\mathcal{C}} + \gamma v'(\mathbf{s}'_k, \mathbf{a}'_k, \mathcal{C}_k)|_{a^{i\prime}=\pi^{i\prime}(s^{\mathcal{C}\prime})}$;
11 Update the characteristic network by minimizing the loss $\mathcal{L}(\theta) = \frac{1}{K}\sum_k (y_k - v(\mathbf{s}_k, \mathbf{a}_k, \mathcal{C}_k))^2$;
12 Calculate the Shapley value as $\phi^i(s,a) := \sum_{\mathcal{C} \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathcal{C}|!(n-|\mathcal{C}|-1)!}{n!} [v(s,a,\mathcal{C} \cup \{i\}) - v(s,a,\mathcal{C})]$;
13 Update actor using the gradient $\nabla_{\theta^i} J \approx \frac{1}{K}\sum_k \nabla_{\theta^i} \pi^i(s^{\mathcal{C}}) \nabla_{a^i} \phi^i(\mathbf{s}_k, \mathbf{a}_k)$ where $a^i = \pi^i(s^{\mathcal{C}})$;
14 **end**
15 Update all target networks: $\theta^{i\prime} \leftarrow \tau\theta^i + (1-\tau)\theta^{i\prime}$.
16 **end**
17 **end**

Based on the Shapley value defined in Definition 4, we propose the Algorithm 1 for each vehicle to learn a cooperative policy. Instead of modeling the action value function, we use a neural network to learn the characteristic function $v(s,a,\mathcal{C})$. This is used to calculate the Shapley value in step 12. Then we use the Shaley value as the reallocated total reward to learn a cooperative policy, as we know the Shaley value reward reallocation is fair and efficient. Moreover, the Shaley value reward reallocation is stable for a convex game. This algorithm adopts centralized training and decentralized execution paradigm that is first proposed in [19]. The effectiveness of this algorithm is evaluated in the following experiment section.

VI. Experiment

In this section, we use CARLA [39], an open-source simulator that supports the development, training, and validation of autonomous driving systems, to validate our proposed method. The host machine adopted in our experiments is a server configured with Intel Core i9-10900X processors and four NVIDIA Quadro RTX 6000 GPUs. Our experiments are performed on Python 3.5.4, GCC7 7.5, openAI gym 0.10.5, numpy 1.14.5, tensorflow 1.8.0, and CUDA 9.0.

We consider a 3-lane freeway scenario with CAVs as shown in Fig. 1. We use Algorithm 1 for CAVs to learn a cooperative policy for behavior planning to decide whether to change or keep lane. For each vehicle, the action set $\mathcal{A}^i$ includes {Keep Lane (KL), Change Left (CL), Change Right (CR), Emergency Stop (ES)}. The stage-wise reward for each vehicle involves their velocity and comfort. The comfort of a vehicle (for passenger's experience) is defined based on its acceleration and action a^i as follows:

$$comfort = \begin{cases} 3, & \text{if } |acceleration| < \Theta \text{ and } a^i = KL; \\ 2, & \text{if } |acceleration| \geq \Theta \text{ and } a^i = KL; \\ 1, & \text{if } a^i = CL/CR; \\ 0, & \text{if in } ES. \end{cases} \tag{9}$$

where Θ is a predefined threshold. The reward function for vehicle i is defined as:

$$R(s,a,\{i\}) = w \cdot velocity + comfort, \tag{10}$$

where w is a trade-off weight. We randomly sample a coalition $\mathcal{C}$ at each timestep. All vehicles in the coalition $\mathcal{C}$ communicate and cooperate with other coalition members. Let $R(s,a,\mathcal{C}) = \sum_{i \in \mathcal{C}} R(s,a,\{i\})$ to represent the coalition's state-wise reward. While training, we use Shapley value to reallocate the reward to encourage communication-based cooperation and improve the system's total reward in Eq. 3. As analysed in Section V, our reward reallocation method is fair (Proposition 1), efficient (Theorem 1), and stable (Theorem 2).

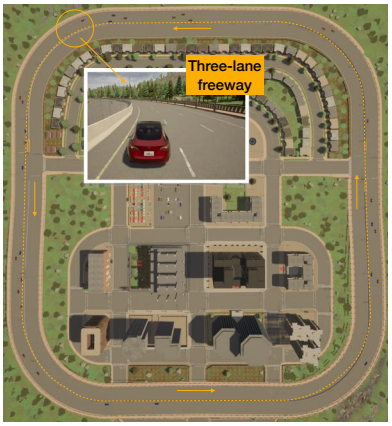

Fig. 1. The example scenario of a 3-lane freeway in CARLA [39]. Vehicles are scattered on the outer-loop of the map "Town 05". The environment can be mixed traffic with both autonomous and human-driven vehicles.

Each vehicle's states include its position, velocity, acceleration, image and point clouds that are captured by the onboard camera and LIDAR sensors respectively. The resolution of camera image is $375 \times 1,242$ pixels. Each point cloud from LIDAR is stored with the 3 coordinates (the ego vehicle being the origin), representing forward, left, and up respectively, and an additional reflectance value. We use methods in [40] to process images and point clouds together, with Ultra-Fast-Lane-Detection [41] (ResNet-18 [42] as the backbone) and PointPillars [43] for 3D object detection.

We assume all CAVs share their states, actions, and environment perception with others under the agreement constructed using Shapley value defined in Definition 4. We use a neural network to learn the characteristic function $v(s, a, \mathcal{C})$ and then use it to calculate the Shapley value as the reallocated total reward. Then we update the local policy by the Shapley value as shown in Alg. 1. We use recent advances in safe RL: model predictive shielding (MPS) that has formal safety guarantee [44], [45] to maintain a safe learning process.

A. Comparison with Baselines

We evaluate Algorithm 1 against the state-of-the-art methods using 30 CAVs to show how Shapley value-based reward reallocation can be used to learn a cooperative policy to improve the system's total reward. We follow the open-source implementations from MADDPG [19], M3DDPG [46], and COMA [47]. For all algorithms, agents are trained for 100k episodes with 5 random seeds and a maximum of 40 steps for each episode. We compare the mean episode system reward (system's total reward averaged over every 1000 episodes) as shown in Fig. 2. We observe that the Shapley value-based reward reallocation outperforms the state-of-the-art methods. This is because our method encourages more cooperation among vehicles. Vehicles can change lanes cooperatively to get a larger average velocity and comfort for the entire system. We also observe that our method converges slower because it relies on the estimation of the Shapley value which is not accurate at the beginning. The M3DDPG is a bit conservative because it considers the robustness to the worst-case scenario.

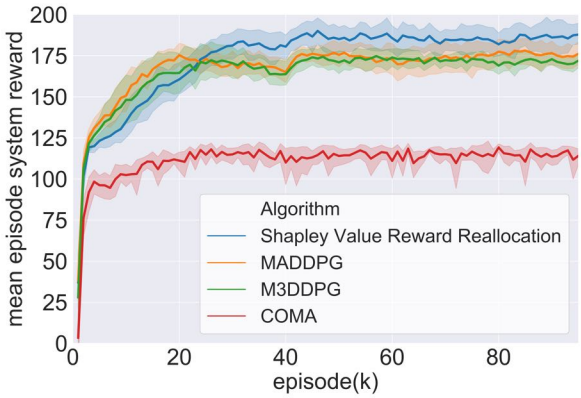

Fig. 2. Comparison between our Shapley value reward reallocation method with baselines. Our method gets larger mean episode system reward.

B. Mixed Traffic

In this section, we show that our algorithm is also effective for CAVs in mixed traffic scenarios where the traffic includes both CAVs and human-driven vehicles. In this set of experiments, the total number of vehicles is 30. We change the CAV ratio (the total CAV number divided by the total number of all vehicles) from 0 to 1 as listed in Table I. We use our Alg. 1 and COMA for CAV's policy learning in two sets of experiments. We use CARLA's built-in autopilot mode to simulate human-driven vehicles [39]. As human-driven vehicle cannot share their states and actions, the CAVs can only form coalitions to estimate the centralized critic using $Q_{\pi^{\mathcal{C}}}(s^{\mathcal{C}}, a^{\mathcal{C}})$ that only has information from agents within coalition $\mathcal{C}$. The joint policy of coalition $\mathcal{C}$ is $\pi^{\mathcal{C}}(a^{\mathcal{C}}|s^{\mathcal{C}})$. In the implementation, we simply set the missing input of the neural network to be zero.

TABLE I

THE SYSTEM EFFICIENCY COMPARISON UNDER THE MIXED TRAFFIC.

CAV ratio	CAV number	human-driven vehicle number	average velocity (mph)	average comfort
Algorithm	Shapley Value Reward Reallocation			
0	0	30	60.06	2.61
0.17	5	25	61.75	2.63
0.33	10	20	64.71	2.68
0.5	15	15	65.12	2.71
0.67	20	10	65.19	2.74
0.83	25	5	65.49	2.76
1	30	0	66.14	2.81
Algorithm	COMA			
0	0	30	60.06	2.61
0.17	5	25	61.19	2.62
0.33	10	20	63.15	2.65
0.5	15	15	63.44	2.68
0.67	20	10	63.48	2.69
0.83	25	5	63.69	2.71
1	30	0	64.12	2.75

We compare the average velocity and comfort for all vehicles under different CAV ratios. The velocity and comfort are averaged over all the 40,000 timesteps used in the simulation. The result in Table I shows the average velocity and comfort of the entire mixed traffic. From the result using Shapley value reward reallocation, we can see the average velocity and comfort increase when the CAV ratio gets higher. Comparing the pure CAVs' case and the pure human-driven vehicles' case, the average velocity improves 10% and the average comfort improves 8%. The results also give us insights that the penetration of the CAVs can improve traffic in the future. Comparing with COMA, our algorithm gets a higher system's total reward in terms of both average velocity and comfort.

VII. CONCLUSION

In this work, we propose a Shapley value-based method to reallocate the system's total reward to each agent, to motivate cooperation among agents, for multi-agent systems such as connected autonomous vehicles (CAVs). We prove that the proposed Shapley value-based reward reallocation locates in the core of the convex game. Hence, the reward reallocation mechanism is stable and efficient, and each individual agent should stay in the cooperating coalition to receive more rewards. We design a cooperative policy learning algorithm which is centralized training and distributed execution. In experiments, we show the improvement of the system's total reward for CAV systems using the proposed algorithm. We also validate the effectiveness of our method in a mixed traffic scenario. In the future, we will leverage the results to other networked CPS to better understand the benefits of communication-based cooperation.

REFERENCES

[1] J. B. Kenney, "Dedicated short-range communications (dsrc) standards in the united states," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1162–1182, July 2011.

[2] J. Rios-Torres and A. A. Malikopoulos, "A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1066–1077, May 2017.

[3] J. Lee and B. Park, "Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 81–90, March 2012.

[4] T. Ort, L. Paull, and D. Rus, "Autonomous vehicle navigation in rural environments without detailed prior maps," in *ICRA 2018*. IEEE, 2018, pp. 2040–2047.

[5] J. Ploeg, D. P. Shukla, N. van de Wouw, and H. Nijmeijer, "Controller synthesis for string stability of vehicle platoons," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 854–865, April 2014.

[6] K. Y. Liang, J. Mårtensson, and K. H. Johansson, "Heavy-duty vehicle platoon formation for fuel efficiency," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1051–1061, April 2016.

[7] S. Darbha, S. Konduri, and P. R. Pagilla, "Benefits of v2v communication for autonomous and connected vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1954–1963, 2019.

[8] M. Bansal, A. Krizhevsky, and A. S. Ogale, "Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst," *CoRR*, vol. abs/1812.03079, 2018.

[9] A. Amini and I. Gilitschenski, "Learning robust control policies for end-to-end autonomous driving from data-driven simulation," *IEEE Robot. Autom. Lett.*, vol. 5, pp. 1–1, 01 2020.

[10] M. Henaff, Y. LeCun, and A. Canziani, "Model-predictive policy learning with uncertainty regularization for driving in dense traffic," in *7th ICLR*, 1 2019.

[11] K. Jang and E. Vinitsky, "Simulation to scaled city: Zero-shot policy transfer for traffic control via autonomous vehicles," in *10th ICCPS*, ser. ICCPS '19, 2019, p. 291–300.

[12] L. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 301–317, 1953.

[13] J. Wang, Y. Zhang, T.-K. Kim, and Y. Gu, "Shapley q-value: A local reward approach to solve global reward games," in *AAAI*, 2020.

[14] K. Zhang and Z. Yang, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv:1911.10635*, 2019.

[15] P. Sunehag and G. Lever, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *AAMAS*, 2018, pp. 2085–2087.

[16] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, "Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning," in *NeurIPS*, December 2020.

[17] J. Foerster and G. Farquhar, "Counterfactual multi-agent policy gradients," in *AAAI*, 2018.

[18] M. Zhou, Z. Liu, P. Sui, Y. Li, and Y. Y. Chung, "Learning implicit credit assignment for cooperative multi-agent reinforcement learning," in *NeurIPS*, vol. 33, 2020, pp. 11 853–11 864.

[19] R. Lowe and Y. I. Wu, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *NeurIPS*, 2017, pp. 6379–6390.

[20] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *ICML*, 2019, pp. 2961–2970.

[21] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," in *NeurIPS*. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 2252–2260.

[22] D. Kim, S. Moon, D. Hostallero, W. J. Kang, T. Lee, K. Son, and Y. Yi, "Learning to schedule communication in multi-agent reinforcement learning," in *ICLR*, 2019.

[23] C. Zhang and V. Lesser, "Coordinated multi-agent reinforcement learning in networked distributed pomdps," in *AAAI*, 2011.

[24] F. Wu, S. Zilberstein, and X. Chen, "Online planning for multi-agent systems with bounded communication," *Artificial Intelligence*, vol. 175, no. 2, pp. 487 – 511, 2011.

[25] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *NeurIPS*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2137–2145.

[26] J. Jiang and Z. Lu, "Learning attentional communication for multi-agent cooperation," in *NeurIPS*. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 7265–7275.

[27] A. Okada, "A noncooperative coalitional bargaining game with random proposers," *Games and Economic Behavior*, vol. 16, no. 1, pp. 97–108, 1996.

[28] B. Moldovanu and E. Winter, "Order independent equilibria," *Games and Economic Behavior*, pp. 21–35, 1995.

[29] I. Vakilinia and S. Sengupta, "Fair and private rewarding in a coalitional game of cybersecurity information sharing," *IET Information Security*, vol. 13, no. 6, pp. 530–540, 2019.

[30] T. W. Sandholm, "Distributed rational decision making," *Multiagent systems: a modern approach to distributed artificial intelligence*, pp. 201–258, 1999.

[31] S. Kraus, O. Shehory, and G. Taase, "The advantages of compromising in coalition formation with incomplete information," in *AAMAS*, vol. 4, 2004, pp. 588–595.

[32] G. Chalkiadakis and C. Boutilier, "Bayesian reinforcement learning for coalition formation under uncertainty," in *AAMAS*, 2004, pp. 1090–1097.

[33] K. Taywade, "Multi-agent reinforcement learning for decentralized coalition formation games," in *AAAI*, vol. 35, no. 18, 2021, pp. 15 738–15 739.

[34] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *ICML*. PMLR, 2019, pp. 2242–2251.

[35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, vol. 30, 2017.

[36] T. P. Michalak, K. V. Aadithya, P. L. Szczepanski, B. Ravindran, and N. R. Jennings, "Efficient computation of the shapley value for game-theoretic network centrality," *J. Artif. Int. Res.*, vol. 46, no. 1, p. 607–650, Jan. 2013.

[37] G. Chalkiadakis, E. Elkind, and M. Wooldridge, "Computational aspects of cooperative game theory," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 6, pp. 1–168, 2011.

[38] S. C. Littlechild and G. Owen, "A simple expression for the shapley value in a special case," *Management Science*, vol. 20, no. 3, pp. 370–372, 1973.

[39] A. Dosovitskiy and G. Ros, "CARLA: An open urban driving simulator," in *CoRL*, 2017, pp. 1–16.

[40] S. Zhou, M. Xie, Y. Jin, F. Miao, and C. Ding, "An end-to-end multi-task object detection using embedded gpu in autonomous driving," in *22nd International Symposium on Quality Electronic Design*. ISQED, 2021, pp. 122–128.

[41] Z. Qin, H. Wang, and X. Li, "Ultra fast structure-aware deep lane detection," *arXiv:2004.11757*, 2020.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[43] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019, pp. 12 697–12 705.

[44] S. Li and O. Bastani, "Robust model predictive shielding for safe reinforcement learning with stochastic dynamics," in *ICRA*, 2020, pp. 7166–7172.

[45] W. Zhang, O. Bastani, and V. Kumar, "Mamps: Safe multi-agent reinforcement learning via model predictive shielding," *arXiv:1910.12639*, 2019.

[46] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and S. Russell, "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient," in *AAAI*, vol. 33, no. 01, 2019, pp. 4213–4220.

[47] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *AAAI*, vol. 32, no. 1, 2018.