


Rollout-based Shapley Values for Explainable Cooperative Multi-Agent Reinforcement Learning

Franco Ruggeri 


Ericsson Research
KTH Royal Institute of Technology
Stockholm, Sweden
franco.ruggeri@ericsson.com

William Emanuelsson


Ericsson Research
KTH Royal Institute of Technology
Stockholm, Sweden
wem@kth.se

Ahmad Terra 

Ericsson Research
KTH Royal Institute of Technology
Stockholm, Sweden
ahmad.terra@ericsson.com

Rafia Inam 

Ericsson Research
KTH Royal Institute of Technology
Stockholm, Sweden
rafia.inam@ericsson.com
raina@kth.se

Karl H. Johansson 

KTH Royal Institute of Technology
Stockholm, Sweden
kallej@kth.se

Abstract—Credit assignment in cooperative Multi-Agent Reinforcement Learning (MARL) focuses on quantifying individual agent contributions toward achieving a shared objective. One widely adopted approach to compute these contributions is through the application of Shapley Values, a concept derived from game theory. Previous research in Explainable Reinforcement Learning (XRL) successfully computed Global Shapley Values (GSVs), albeit neglecting local explanations in specific situations. In contrast, another approach concentrated on learning Local Shapley Values (LSVs) during training, prioritizing sample efficiency over explainability. In this paper, we extend an existing method to generate local and global explanations in a model-agnostic manner, bridging the gap between these two approaches. We apply our proposed algorithm to two cooperative tasks: a predator-prey environment and an antenna tilt optimization problem in cellular networks. Our findings reveal that the LSVs offer valuable insights into the agents' behavior with a finer time-frame granularity, while their aggregation in GSVs enhances trust by potentially identifying suboptimality. Importantly, our approach surpasses the existing state-of-the-art methods in estimating LSVs, enhancing the accuracy of assessing individual agent contributions. This work represents a significant advancement in the field of XRL and provides a powerful tool for gaining deeper insights into agents' behavior in cooperative MARL systems.

Index Terms—Explainable Reinforcement Learning, Multi-Agent Reinforcement Learning, Shapley Values, Remote Electrical Tilt Optimization

I. INTRODUCTION

In recent years, Reinforcement Learning (RL) has emerged as a powerful paradigm for training agents to make sequential decisions in complex, dynamic environments. From autonomous robotics to telecommunication systems, RL has shown remarkable promise in a various applications. However, as RL systems become increasingly integrated into real-world

applications, there arises an urgent and compelling need for transparency and interpretability, especially when the decisions impact people's lives and society.

Explainable Reinforcement Learning (XRL) refers to the ability to elucidate how RL agents make decisions in a way that is easily understandable to human stakeholders [1]. Explanations can be either *local*, focusing on individual decisions, or *global*, encompassing the holistic behavior of the agents. Furthermore, state-of-the-art XRL includes *model-agnostic* techniques, universally applicable to any RL agent and typically extracting explanations after training (*post-hoc*), and *intrinsic* (or model-specific) methods, strictly coupled to the underlying learning algorithm.

The demand for explainability holds considerable importance in the domain of Multi-Agent Reinforcement Learning (MARL). Cooperative MARL explores scenarios where multiple agents collaborate to achieve shared goals [2]. In such settings, discerning the individual contributions of each agent, hereafter referred to as the *credit assignment* problem, presents a challenging task, given that agents exert influence not only directly on the system but also on each other. Nonetheless, illuminating the roles of individual agents within MARL systems is crucial for optimal performance, fairness, reliability, and accountability, given that agent contributions can pinpoint sources of errors or suboptimal behavior [3].

Shapley Values (SVs) [4], rooted in cooperative game theory, have gained recognition as a valuable tool for fairly distributing a payoff among a group of players. Previous research has applied SVs to solve the credit assignment problem in cooperative MARL. Wang et al. [5] devised Shapley Q-Value Deep Deterministic Policy Gradient (SQDDPG), an MARL algorithm that leverages SVs to distribute a global shared reward among the agents for each state and joint action, to improve sample efficiency. However, the generated learning-based SVs cannot be reliably used as local explanations, as

This work was partially supported by the Swedish Foundation for Strategic Research and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

they are accurate only upon optimal training and, therefore, cannot identify problems such as suboptimality. Heuillet et al. [3] introduced a model-agnostic, post-hoc algorithm employing Monte Carlo simulations to approximate SVs, designed for delivering global explanations. Nonetheless, a notable limitation of their approach lies in its inability to provide local explanations, a point highlighted by the authors themselves.

This paper extends the application of SVs in MARL to offer a comprehensive solution that provides both local and global explanations of agent contributions. We propose a modification of [3] for estimating SVs as a function of the state in a model-agnostic, post-hoc manner, achieving the advantages of [3], [5] while overcoming their limitations. Our experimental investigation applies our algorithm to a well-established benchmark task, the *predator-prey* problem, commonly used in prior MARL research. Moreover, we explore uncharted territory by conducting experiments on *Remote Electrical Tilt (RET) optimization*, a prevalent problem in the telecommunication domain [6]. Our findings reveal that local explanations help to explain the internal agents' behavior, offering enhanced temporal granularity. Additionally, global explanations improve trust by summarizing the importance of each agent and identifying suboptimalities. These combined factors demonstrate the effectiveness of our approach and its applicability across diverse domains.

II. RELATED WORK

The literature relevant to this paper encompasses SVs, XRL, and the intersection between them. SVs [4] were initially devised to assess the value of individual players in contributing to the overall payoff. This concept subsequently found widespread application across a diverse range of fields. Lundberg et al. [7] applied SVs to generic machine learning models, considering each input feature as a player. The resulting method, known as SHapley Additive exPlanations (SHAP), calculates the impact of each feature for an individual prediction compared to a baseline prediction. SHAP was later adapted for single-agent RL policies to compute the importance of each state feature to predict an action [8].

Previous work on XRL includes several promising techniques, not only based SVs [1]. Liu et al. [9] presented Linear Model U-Trees, transparent tree-structured RL models whose leaves are equipped with linear models, to mimic complex models following the concept of knowledge distillation. Madumal et al. [10] proposed to build a causal graph to represent the influence of actions on states and rewards. Such a graph could then extract *why* and *why-not* explanations. Hayes et al. [11] designed a framework known as autonomous policy explanation to summarize a policy using minimal logical clauses of user-interpretable abstractions of state features. In environments where the reward is designed as a sum of different components, reward decomposition [12] can provide contrastive explanations to compare two actions in terms of expected cumulative reward components. Terra et al. [13] combined ideas from reward decomposition [12] and SHAP [7] to generate detailed explanations, as well as identify and mitigate

bias. While the works mentioned above focus on a single agent to explain correlations between observations and actions, this paper proposes an XRL method that focuses on the holistic view of a MARL system to find agent contributions. Our approach complements other algorithms that might still be applied to the individual agents in the MARL system.

In cooperative MARL, SVs can be utilized by considering each agent as a player. Wang et al. [5] used SVs to fairly distribute the global reward among the agents, aiming to improve the sample efficiency compared to using the global reward. In this algorithm, named SQDDPG, the SVs are learned during training by approximating with a neural network, for each state and action, the contribution of each agent to each coalition. However, learning-based SVs cannot be reliably exploited as local explanations, as their accuracy is coupled to policy optimality. Thus, suboptimal SQDDPG agents would provide inaccurate SVs that cannot be utilized to identify problems and improve the agents. In this paper, we compare learning-based SVs generated by SQDDPG and rollout-based SVs generated by our algorithm against exact values. A different approach, focused on explainability, was used by Heuillet et al. [3], who approximated the global contribution of each agent employing Monte Carlo simulations. However, the main limitation of such an approach is that it only provides a global value for each agent, neglecting local contributions in specific situations. In this paper, we overcome such a limitation by adapting the algorithm to provide local and global contributions.

III. BACKGROUND

In this section, we outline the mathematical foundations of MARL and SVs, which serve as the basis for our algorithm.

A. Multi-Agent Reinforcement Learning

In MARL, a group of agents interacts within a shared environment, aiming to maximize their respective cumulative rewards. A MARL problem can be formally modeled as a stochastic game known as *Markov game* [14]. A Markov game is defined by a tuple $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i \in \mathcal{N}}, \{R^i\}_{i \in \mathcal{N}}, \mathcal{T}, \gamma)$, where $\mathcal{N} = \{1, \dots, n\}$ is the set of $n > 1$ agents, \mathcal{S} is the state space of the environment, and \mathcal{A}^i is the action space of agent $i \in \mathcal{N}$. Let $\mathcal{A} = \mathcal{A}^1 \times \dots \times \mathcal{A}^n$, then $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the transition probability from any state $s \in \mathcal{S}$ to any state $s' \in \mathcal{S}$ for any joint action $a \in \mathcal{A}$; $R^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward function that determines the immediate reward $r^i = R^i(s, a, s')$ received by agent i for a transition from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ after the joint action $a \in \mathcal{A}$; $\gamma \in [0, 1)$ is the discount factor. In MARL, transition probabilities \mathcal{T} and reward functions $\{R^i\}_{i \in \mathcal{N}}$ are assumed to be unknown. The objective of each agent i is to maximize the expected discounted return $\mathbb{E} [\sum_{t=0}^{\infty} \gamma^t r_t^i]$.

This work focuses on *fully cooperative* Markov games with *terminal state*. In fully cooperative settings, the agents must collaborate to maximize a shared global reward, namely, $R = R^1 = R^2 = \dots = R^N$. However, the individual contribution of each agent to collect the global reward is not present, which raises the problem of *credit assignment*. Having

a terminal state guarantees that the game terminates in a finite time, almost surely.

B. Shapley Values

SV [4] is one of the most popular methods to solve the credit assignment problem in cooperative games. Formally, a *cooperative game* is defined as a tuple (\mathcal{N}, v) , where \mathcal{N} is the set of players and $v : 2^{\mathcal{N}} \rightarrow \mathbb{R}$ is the payoff function that maps a coalition $\mathcal{C} \subseteq \mathcal{N}$ (i.e., a subset of players) to a value $v(\mathcal{C})$, with $v(\emptyset) = 0$. Then, the SV of player $i \in \mathcal{N}$ is defined as:

$$\phi_i = \frac{1}{|\mathcal{N}|} \sum_{\mathcal{C} \subseteq \mathcal{N} \setminus \{i\}} \binom{|\mathcal{N}|-1}{|\mathcal{C}|} v(\mathcal{C} \cup \{i\}) - v(\mathcal{C}) \quad (1)$$

Intuitively, a SV calculates the average of the player's marginal contribution to all possible coalitions to capture not only the player's direct impact on the game but also the interactions with the other agents.

The exact computation of Eq. (1) has an exponential complexity $\mathcal{O}(2^{|\mathcal{N}|-1})$ and, for stochastic games, requires to replace payoff values with expectations, which are generally unknown. Therefore, SVs generally need to be approximated. A possible solution is to use Monte Carlo sampling [3]:

$$\phi_i \approx \hat{\phi}_i = \frac{1}{M} \sum_{k=1}^M [v(\mathcal{C}_k \cup \{i\}) - v(\mathcal{C}_k)], \mathcal{C}_k \sim p_i(\cdot) \quad (2)$$

where $p_i(\mathcal{C}) = \frac{1}{|\mathcal{N}|} \binom{|\mathcal{N}|-1}{|\mathcal{C}|}^{-1}$, $\forall \mathcal{C} \subseteq \mathcal{N} \setminus \{i\}$ is the probability of coalition \mathcal{C} and corresponds to the same weight assigned to coalition \mathcal{C} in Eq. (1). The number of sampled coalitions $M > 0$ offers a trade-off between accuracy and efficiency. A larger M means a better SV approximation but comes at the expense of more computation (the complexity is $\mathcal{O}(M)$).

IV. ROLLOUT-BASED SHAPLEY VALUES

In this section, we define the concepts of Local Shapley Values (LSVs) and Global Shapley Values (GSVs) and present the modifications we made to the algorithm in [3].

A. Shapley Values in Multi-Agent Reinforcement Learning

In fully cooperative Markov games, SVs provide the individual contribution of each agent towards the shared global reward. Let $\pi_{\mathcal{C}}$ be the policy of a coalition $\mathcal{C} \subseteq \mathcal{N}$ where only the agents in the coalition are allowed to take action. We define a LSV $\phi_i^l(s)$ by plugging into Eq. (1) the non-discounted value function as payoff function:

$$v^l(s, \mathcal{C}) = \mathbb{E} \left[\sum_{t=0}^{\infty} r_t \middle| s_0 = s, \pi_{\mathcal{C}} \right] \quad (3)$$

Intuitively, a LSV represents the expected contribution of agent $i \in \mathcal{N}$ starting from the *specific state* $s \in \mathcal{S}$.

Similarly, we define a GSV ϕ_i^g by using the expected return as payoff function in Eq. (1):

$$v^g(\mathcal{C}) = \mathbb{E} \left[\sum_{t=0}^{\infty} r_t \middle| \pi_{\mathcal{C}} \right] \quad (4)$$

Algorithm 1 Rollout-based LSV estimation

Require: Agent i ; State: s ; Policy π ; Episode length T ; Number of rollouts M ;

Ensure: Marginal contribution: δ^i

```

1: for  $1 \leq m \leq M$  do
2:    $\mathcal{C}_{-i} \leftarrow \text{sample\_coalition\_without}(i)$ 
3:    $\mathcal{C}_{+i} \leftarrow \mathcal{C}_{-i} \cup \{i\}$ 
4:    $\pi_{-i} \leftarrow \text{exclude\_agents}(\pi, \mathcal{C}_{-i})$ 
5:    $\pi_{+i} \leftarrow \text{exclude\_agents}(\pi, \mathcal{C}_{+i})$ 
6:    $\text{reset\_environment\_and\_seed}(s, m)$ 
7:    $R_{-i} \leftarrow \text{rollout\_environment}(\pi_{-i})$ 
8:    $\text{reset\_environment\_and\_seed}(s, m)$ 
9:    $R_{+i} \leftarrow \text{rollout\_environment}(\pi_{+i})$ 
10:   $\hat{\phi}_i \leftarrow \hat{\phi}_i + (R_{+i} - R_{-i})$ 
11: end for
12:  $\hat{\phi}_i \leftarrow \hat{\phi}_i / M$ 

```

where, unlike Eq. (3), the initial state s_0 is an additional random variable in the expectation. Intuitively, a GSV represents the expected *overall* contribution of agent $i \in \mathcal{N}$ in the environment. This definition was used also in [3]. The assumption of Markov game with terminal state (see Section III-A) guarantees $v^l(s, \mathcal{C}) < \infty$ and $v^g(\mathcal{C}) < \infty$.

B. Rollout-based estimation

We propose a rollout-based¹ algorithm, adapted from [3], to compute the Monte Carlo approximation of LSVs based on Eqs. (2) and (3). The algorithm, whose pseudo-code is shown in Algorithm 1, estimates the LSV $\hat{\phi}_i(s)$ by computing M marginal contributions $R_{+i} - R_{-i}$ of agent i to randomly sampled coalitions \mathcal{C}_{-i} . The main modifications from [3] to achieve *locality* are (lines 6 and 8): (i) each rollout starts from the state of interest s ; (ii) each marginal contribution $R_{+i} - R_{-i}$ is computed from two rollouts with the same seed, thereby isolating the contribution of agent i to the coalition \mathcal{C}_{-i} from the environment randomness. The `exclude_agents()` function consists of constraining the agents *not in* the coalition $i \notin \mathcal{C}$ to select a *no-operation* action, which in [3] was empirically found to be the best approach to exclude agents from a coalition.

GSVs can be computed by averaging LSVs. In this paper, we average over LSVs at each time step of multiple episodes. Specifically, given an episode, we compute and collect the LSVs at each time step along the episode. The exploration of alternatives, such as using only the LSVs of the initial state of each episode, is left for future work.

Similar to [3], our approach requires full control of the environment to perform rollouts and reset to specific states. This requirement can be satisfied through simulators or, where not available, by learning the system dynamics as commonly done in model-based RL [15]. The computational complexity of the LSV estimation is $\mathcal{O}(M)$, with M controlling the trade-off between accuracy and computation time. Parallelization is

¹In this context, rollout refers to simulating an episode until termination to record the episodic return.

TABLE I: Hyperparameters used to train SQDDPG agents.

Training episodes	5000
Discount factor	0.99
Hidden layers for actor & critic	1
Hidden units for actor & critic	128
Learning rate for actor	1e-4
Learning rate for critic	5e-4
Update frequency of behavior network	100
Update frequency for target network	200
Soft target update coefficient	0.1
Entropy regularization coefficient	1e-3
Batch size	128

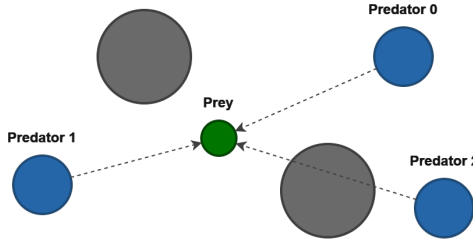


Fig. 1: Illustration of the predator-prey environment with three blue predators attempting to capture the single green prey, while avoiding two grey obstacles.

possible, as each rollout can be run independently in parallel with the others.

V. EXPERIMENTS AND RESULTS

In this section, we describe the experimental setup and present the results of our experiments. These experiments include the empirical evaluation of the LSV accuracy and the qualitative assessment of the local and global explanations provided by LSVs and GSVs, respectively.

A. Experimental setup

All RL agents were trained using SQDDPG [5]. The hyperparameters are reported in Table I. The experiments were conducted using the following two MARL environments:

1) *Predator-prey* [3], [5] : Predators must catch a set of preys in a two-dimensional world while avoiding fixed obstacles. Each predator observes its position, velocity, relative positions to other agents and preys, and preys' velocities. In addition, each predator has five possible actions: accelerate up, down, left, right, and stop (modified to be an acceleration in the opposite direction to the current velocity). The *no-operation* action corresponds to the stop action. We used 3 predators, 1 prey, and 2 obstacles, as illustrated in Fig. 1. The prey moves randomly, whereas the predators are RL agents. The global reward includes the negative distance between the prey and the predator closest to it and a positive reward if the prey is caught. An episode terminates when the prey is caught or after 200 time steps. We made the environment *deterministic*, meaning that the prey always follows the same trajectory given an initial environment state.

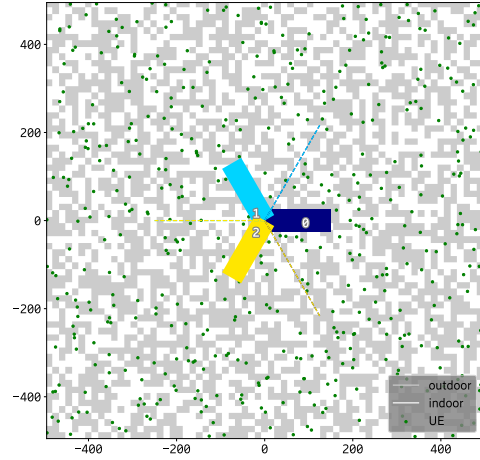


Fig. 2: Illustration of the RET environment with three antennas, indoor and outdoor space, and a uniform random distribution of 500 UEs. Other UE distributions were also used, as described in the experiments.

TABLE II: Configuration of the RET environment.

Map size	10 000 m ²
Number of BSs	1
Range of tilt angle	[0°, 15°]
Downlink maximum power	40 W
Antenna model on BS	hv 742215 fitted
Antenna type on UEs	Isotropic
Propagation model	Okumura-Hata
Fraction of indoor/outdoor	0.5
Radio technology	LTE

2) *Remote Electrical Tilt (RET)* [6]: A set of Base Stations (BSs) and User Equipments (UEs) are deployed in a cellular network. Each BS is equipped with 3 directional antennas. The vertical electrical tilt of each antenna needs to be adjusted to optimize a network Key Performance Indicator (KPI). We used 1 BS and several UE distributions depending on the specific experiment, as described in the following sub-sections. Each antenna observes the average Signal-to-Interference-plus-Noise Ratio (SINR) and Reference Signal Received Power (RSRP) measured from the set of UEs it serves, as well as its tilt angle. In addition, each antenna can adjust its tilt angle incrementally by 1° or keep the same tilt angle. The *no-operation* action, used for the rollout-based LSV estimation, consists of setting the antenna's transmission power to 0 W and does not belong to the action space. The global reward is the percentage of UEs in the network with $\text{SINR} \geq 3\text{ dB}$ (i.e., good signal quality) and $\text{RSRP} \geq -108\text{ dBm}$ (i.e., good coverage). An episode terminates after 20 time steps. The experiments were conducted using a system simulation tool that implements a simplified map-based ray-tracing propagation for arriving at the path-gains at various UE drops [16]. Table II contains the configuration parameters, whereas an illustration of the environment with a uniform random distribution of UEs is shown in Fig. 2.

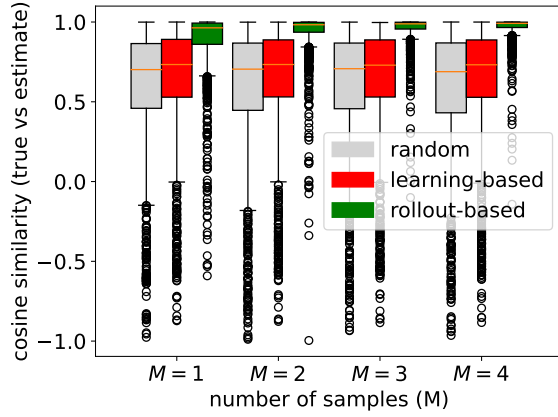


Fig. 3: Cosine similarity between the true and estimated LSVs using rollout-based, learning-based, and random estimation, for various numbers of samples M in the predator-prey environment. The statistics are calculated across 1000 episodes.

B. Accuracy evaluation

We evaluated the accuracy of LSV estimated by our rollout-based algorithm in the deterministic predator-prey environment. This environment, being deterministic, allowed us to precisely compute the payoff function (Eq. (3)) with a single rollout². We computed exact LSVs by examining all eight possible predator coalitions and compared them to estimates obtained via SQDDPG (learning-based) and via our rollout-based approach, varying the number of sampled coalitions (M). We also included a baseline with random values drawn uniformly between 0 and 1.

As shown in Fig. 3, our algorithm consistently outperformed SQDDPG. The precision of our method improved with higher values of M , while SQDDPG estimates only marginally outperformed random values and showed no notable improvement with increased M . This result aligns with findings from [5], which reported limited performance gains from increasing M in SQDDPG. Several factors can be attributed to the inferior accuracy of learning-based LSVs. First, learning-based LSVs consider a discount factor, causing them to estimate discounted cumulative rewards rather than episodic returns. Nevertheless, the performance of RL agents is typically evaluated based on *non-discounted* episodic returns, and including the discount factor in learning algorithms primarily serves convergence guarantees. Second, learning-based LSVs are susceptible to suboptimal training, meaning that the LSVs accuracy is highly correlated to the agents' performance. In contrast, post-hoc rollout-based estimates can accurately approximate true LSVs independently of the underlying learned policies. This property is highly desirable in explainability, as one of the critical use cases is identifying suboptimal policies (e.g., due to undertraining). These results highlight the superiority of our rollout-based algorithm over SQDDPG for estimating LSVs.

²In the RET environment, due to stochastic procedures in the simulator, computing the exact LSVs is not possible.

C. Local explanations

We experimented with the rollout-based LSVs to qualitatively check whether and how they can be helpful as local explanations in predator-prey and RET environments. For this purpose, given a complete episode, we computed the LSVs at each time step to show their evolution. Fig. 4 shows the trajectories of predators and prey during a complete episode, along with the LSVs and their L1 normalization. In the starting position ($t = 0$), although all the predators are far from the prey, predator 1 is the closest, and its slight advantage is captured by the LSVs. As the episode continues ($0 < t \leq 5$), the prey moves downwards, decreasing the y-distance from predators 0 and 2, while running away from predator 1. Thus, predators 0 and 2 gain importance relative to predator 1, as reflected by the L1-normalized LSVs. From this point ($t > 5$), predator 0 starts losing importance (decreasing LSV), as the others are the candidates to catch the prey. An interesting insight, not immediately visible from the trajectories, is that from $t = 5$ predator 0 has nearly no impact for the rest game ($LSV \approx 0$). Such intuition might be instrumental, for example, to stop agents when they become superfluous, with benefits such as energy savings. Furthermore, just before the episode ends ($t = 14$), the LSVs of predators 1 and 2 drop significantly, as they are so close to the prey that a no-operation action from a predator would still result in a quick catch from the other one or with the prey bumping into a predator. Regardless, the normalized LSVs of predator 2 is slightly higher since its next move will catch the prey.

For the experiment with RET, we used the uniform UE distribution illustrated in Fig. 2, and the results are presented in Fig. 5. The episode starts with random tilt angles for each antenna, as shown in Fig. 5a (antenna 0, 1, and 2 have a tilt angle of 0° , 15° , and 2° , respectively). While antenna 1 starts fully down-tilted and takes the uptilt action until it reached the optimal position (4°), antennas 0 and 2 do the opposite by tilting down. We can see from Fig. 5b that the KPI of each antenna increases over time, and antenna 2 has the highest percentage of well-served UEs. However, the LSVs distribution has a different trend, with antenna 0 having the highest contribution (Figs. 5c and 5d). The LSVs decrease over time because the episode is artificially truncated after 20 time steps, reducing the horizon available to accumulate positive rewards as the episode evolves. In the normalized LSVs (Fig. 5d), the contributions of antenna 0 and 2 decrease at early time steps because of the normalization process, where in each time step, all the contributions must sum up to 1. More importantly, the normalized LSVs converge towards stable relative contributions of the antennas that mirror the underlying UE distribution. Indeed, antenna 0 covers a larger area than the other antennas, given that the simulated area is square and inevitably presents an unbalanced three-zone division (see Fig. 2). This observation aligns with the logic of the *no-operation* action: when an antenna is excluded from the coalition, it is turned off, causing its UEs to be re-assigned to other antennas, deteriorating their KPI. Consequently, the

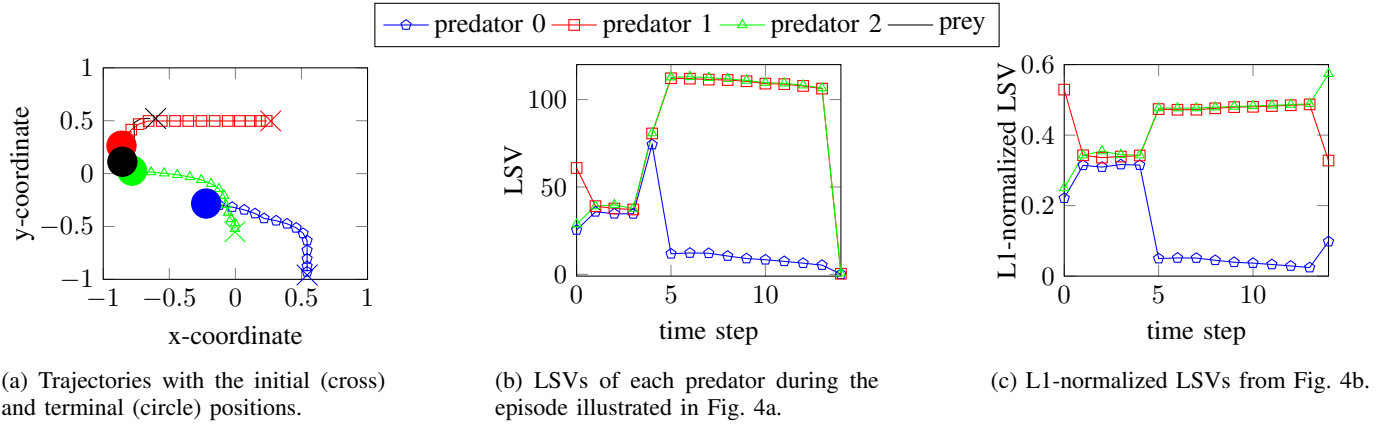


Fig. 4: Trajectories of predators and prey, along with LSVs and their L1-normalized values, for a full episode.

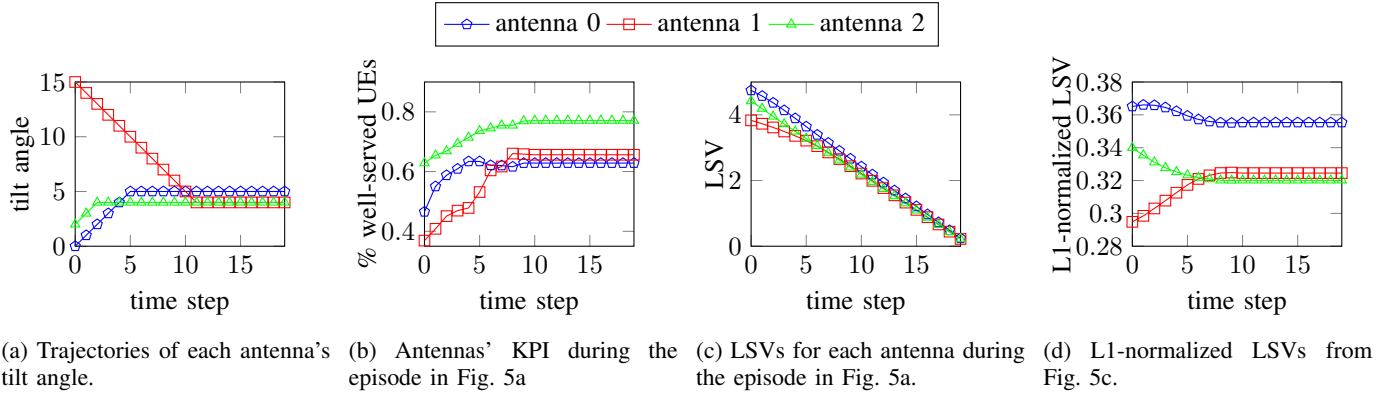


Fig. 5: Trajectories, KPI, LSVs and L1-normalized LSVs for a full episode of the RET environment illustrated in Fig. 2.

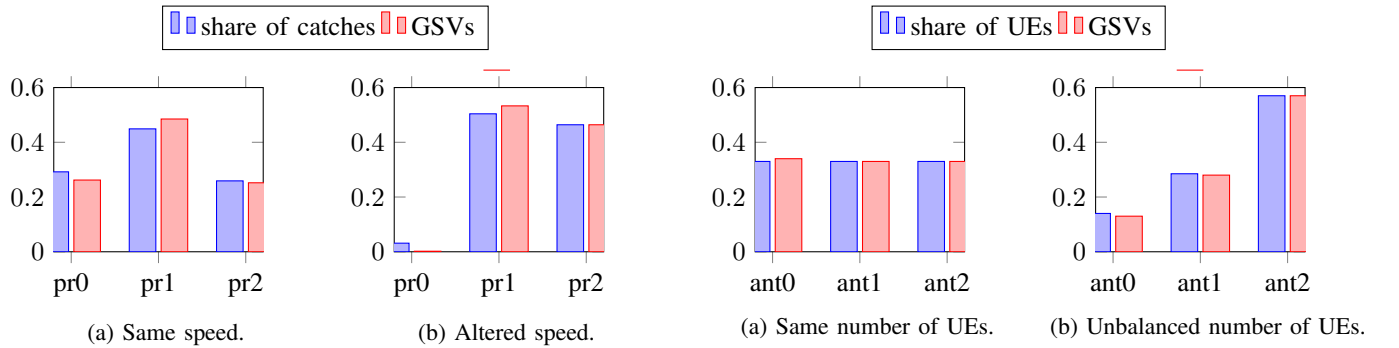


Fig. 6: Share of catches and GSVs for two settings of the predator-prey environment: a) all the predators have the same maximum speed of 1; and b) predator (pr) 0, 1, and 2 with the maximum speed of 0.2, 0.8, and 2.0, respectively.

Fig. 7: Share of UEs and GSVs for two settings of the RET environment: a) all antennas serving 20 symmetrically distributed UEs and b) antenna (ant) 0, 1, and 2 serving 4, 8, and 16 UEs, respectively.

more UEs the antenna serves, the greater impact removing it from the coalition has.

D. Global explanations

We designed experiments to qualitatively assess whether GSVs can correctly capture the overall contribution of each agent and provide insights to human users. In these experi-

ments, we used $M = 1$ because the aggregation of numerous LSVs guarantees to consider each coalition multiple times. In predator-prey, inspired by [3], we computed the GSVs in the original environment, where all the predators have the same maximum speed, and in an altered environment, where the three predators have different maximum speeds (predators 1, 2, and 3 have maximum relative speeds of 0.2, 0.8, and

2.0, respectively). Fig. 6 shows the comparison of the GSVs and the share of prey catches for each predator, a reasonable indicator of the true contributions. In both settings, the GSVs have a distribution very similar to the share of catches (cosine similarity 0.9972 and 0.9987 for the same-speed and altered-speed cases, respectively). In the same-speed scenario, predator 1 has the lead, most likely due to suboptimal or biased policies (e.g., predator 1 responsible for catching the prey, and other predators responsible for patrolling). For the altered-speed scenario, interestingly, the shares of catches do not reflect the relative maximum speeds of the predators. Possible reasons are hidden benefits of having a capped lower speed (e.g., the predator can speed up faster to reach its maximum speed, and its maximum speed is the perfect spot to have good maneuverability) or, again, suboptimal policies. In both cases, GSVs can highlight such unintuitive and unexpected agent contributions, calling the developers for further investigation and improvement.

Similarly, in the RET environment, we computed the GSVs in two controlled scenarios where the set of UEs is deployed in a circumference around the BS (at distance 400 m from the BS and linearly spaced in the intervals $[-\pi/12, \pi/12]$, $[7\pi/12, 9\pi/12]$, and $[-9\pi/12, -7\pi/12]$) in order to minimize the impact of distance and randomness on the agent contributions. The first scenario has the same number of UEs (20) in each sector, whereas the second one has an unbalanced UE distribution (sectors 0, 1, and 2 have 4, 8, and 16 UEs, respectively). Fig. 7 compares the GSVs to the share of UEs, which is a reasonable indicator of the true contributions, as the reward considers the percentage of well-served UEs. In both scenarios, the GSVs have a distribution very similar to the share of UEs (cosine similarity 0.9995 and 0.9997 for the balanced and unbalanced cases, respectively), providing a good indication of the true contribution of each antenna. Unlike the predator-prey experiment, the agent contributions (and therefore the GSVs) match the prior belief and confirm unbiased and well-performing policies.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a rollout-based algorithm to approximate the contribution of individual agents within a MARL system using Shapley Value. Our algorithm provides both local and global explanations. Local explanations highlight the contribution of each agent in specific situations, and our approach has consistently demonstrated superior accuracy compared to SQDDPG. Consequently, more accurate local explanations help to detect biased and suboptimal policies. For global explanations, our algorithm identifies crucial environmental factors and policy suboptimalities, enhancing explainability, accountability, trust, and overall system reliability. An essential requirement of our approach is full control over the system, which is necessary for enabling the execution of rollouts and resets. While we experimented with simulations in this study, real-world applications may necessitate incorporating an additional model to approximate the system dynamics, as commonly done in model-based RL.

Future research may involve evaluating our algorithm with learned system dynamics and experimenting with more intricate scenarios within the RET problem. These scenarios might encompass multiple BSs, potentially introducing interference among them, thus raising additional challenges and needs for local and global insights.

REFERENCES

- [1] E. Puiutta and E. M. S. P. Veith, "Explainable Reinforcement Learning: A Survey," in *Machine Learning and Knowledge Extraction* (A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, eds.), vol. 12279, pp. 77–95, Cham: Springer International Publishing, 2020.
- [2] R. Lowe, YI. WU, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [3] A. Heuillet, F. Couthouis, and N. Diaz-Rodriguez, "Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning With Shapley Values," *IEEE Computational Intelligence Magazine*, vol. 17, pp. 59–71, Feb. 2022.
- [4] L. S. Shapley, "A value for n-Person games," in *Contributions to the Theory of Games II* (H. W. Kuhn and A. W. Tucker, eds.), pp. 307–317, Princeton: Princeton University Press, 1953.
- [5] J. Wang, Y. Zhang, T.-K. Kim, and Y. Gu, "Shapley Q-Value: A Local Reward Approach to Solve Global Reward Games," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7285–7292, Apr. 2020.
- [6] A. Mendo, J. Outes-Carnero, Y. Ng-Molina, and J. Ramiro-Moreno, "Multi-Agent Reinforcement Learning with Common Policy for Antenna Tilt Optimization," *Journal of Computer Science*, vol. 50, no. 3, pp. 883–889, 2023.
- [7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [8] R. Liessner, J. Dohmen, and M. Wiering, "Explainable Reinforcement Learning for Longitudinal Control," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, (Online Streaming, — Select a Country —), pp. 874–881, SCITEPRESS - Science and Technology Publications, 2021.
- [9] G. Liu, O. Schulte, W. Zhu, and Q. Li, "Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees," in *Machine Learning and Knowledge Discovery in Databases* (M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Iffrim, eds.), vol. 11052, pp. 414–429, Cham: Springer International Publishing, 2019.
- [10] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable Reinforcement Learning through a Causal Lens," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 2493–2500, Apr. 2020.
- [11] B. Hayes and J. A. Shah, "Improving Robot Controller Transparency Through Autonomous Policy Explanation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, (Vienna Austria), pp. 303–312, ACM, Mar. 2017.
- [12] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, "Explainable reinforcement learning via reward decomposition," in *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*, 2019.
- [13] A. Terra, R. Inam, and E. Fersman, "BEERL: Both Ends Explanations for Reinforcement Learning," *Applied Sciences*, vol. 12, p. 10947, Oct. 2022.
- [14] K. Zhang, Z. Yang, and T. Başar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," in *Handbook of Reinforcement Learning and Control* (K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, eds.), vol. 325, pp. 321–384, Cham: Springer International Publishing, 2021.
- [15] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker, "Model-based Reinforcement Learning: A Survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 1, pp. 1–118, 2023.
- [16] H. Asplund, M. Johansson, M. Lundevall, and N. Jaldén, "A set of propagation models for site-specific predictions," in *12th European Conference on Antennas and Propagation (EuCAP 2018)*, pp. 1–5, 2018.