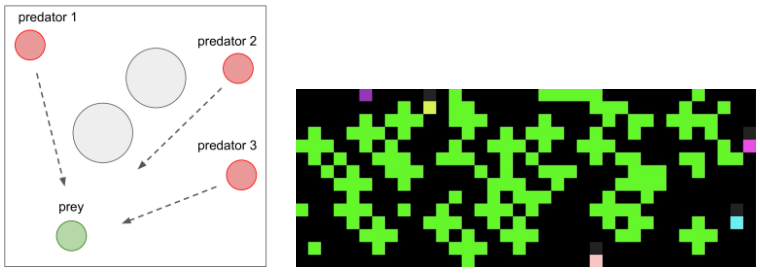
Supplementary Material of Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning with Shapley Values

Alexandre Heuillet, Fabien Couthouis, Natalia Díaz-Rodríguez

2021

Environments



(a) (b)

Figure 1: (a) Predator-Prey scenario of Multiagent Particle environment [1]. (b) Screenshot of the *Harvest* scenario of Sequential Social Dilemmas environment [4]. Green blocks represent apples, while colored blocks are agents whose goal is to collect apples without tarnishing their source. Consecutive chains of apples are trees whose role is to regrow apples at a rate that depends on the number of nearby apples.

Additional results

Extra experiments were conducted to test the scalability of models in Experiment 2 (see Subsection V.C of the article) by increasing the number of agents to 9 predators and 3 preys. However, running over 500,000 episodes MADDPG did not make the predators global reward converge. When observing the model, contrary to the one used in Experiment 1 (see Subsection V.B of the article)

agents do not show any collective cooperative strategy and wander aimlessly on the map failing to catch the prey (except when randomly hitting them). It can be hypothesized that the high number of agents makes cooperation between them difficult and this may be pointed out as a limitation of MADDPG.

We also tried to use Value-Decomposition Networks (VDN) [3] to train a model for the Harvest [4] environment but this was not successful since the model struggled to converge and did not converged to a positive mean reward while A3C reached around 500 of mean reward, see Fig. 7).

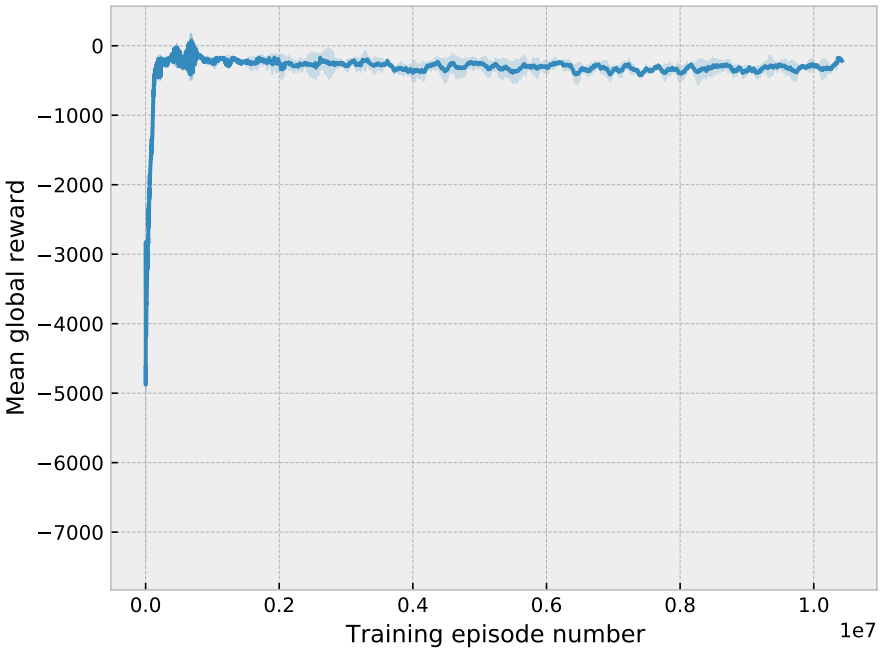


Figure 2: Learning curve of the Harvest VDN model (averaged over 3 runs). It is clear that the model did not reach a way lower max reward than the A3C model in Fig. 7

Social Metrics Per Agent

We further studied the social metrics presented in Subsection V.D.4 of the article and introduced in [2]. They were refactored into a per agent basis in order to make a more relevant comparison with Shapley values. Considering agent i among N agents, the following formulas were obtained:

$$U_i = \mathbb{E}\left[\frac{R_i}{T}\right] \tag{1}$$

$$S_i = t_i = \mathbb{E}[t/r_t^i > 0] \tag{2}$$

$$E_i = 1 - \frac{\sum_{j=1}^{N} |R_i - R_j|}{2\sum_{j=1}^{N} R_j} \tag{3}$$

Where R_i is the reward obtained by agent i. Using these refactored formulas, the metrics were plotted using the same data than in Subsection V.D.4 of the article in Fig. 3, Fig. 4 and Fig. 5.

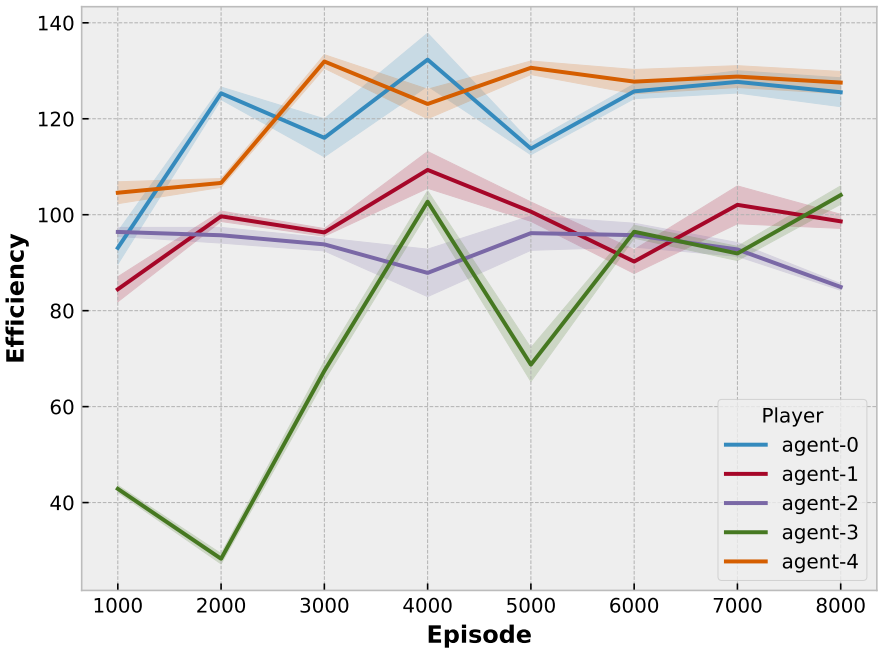


Figure 3: Evolution of *Efficiency* social metric per agent over several episodes (averaged on 4 runs).

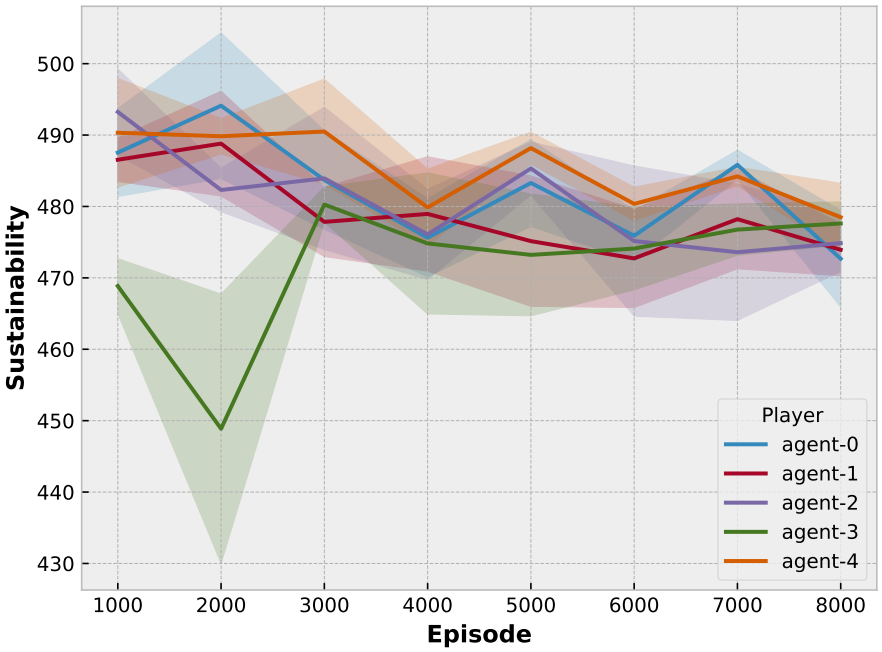


Figure 4: Evolution of Sustainability per agent over several episodes (averaged on 4 runs).

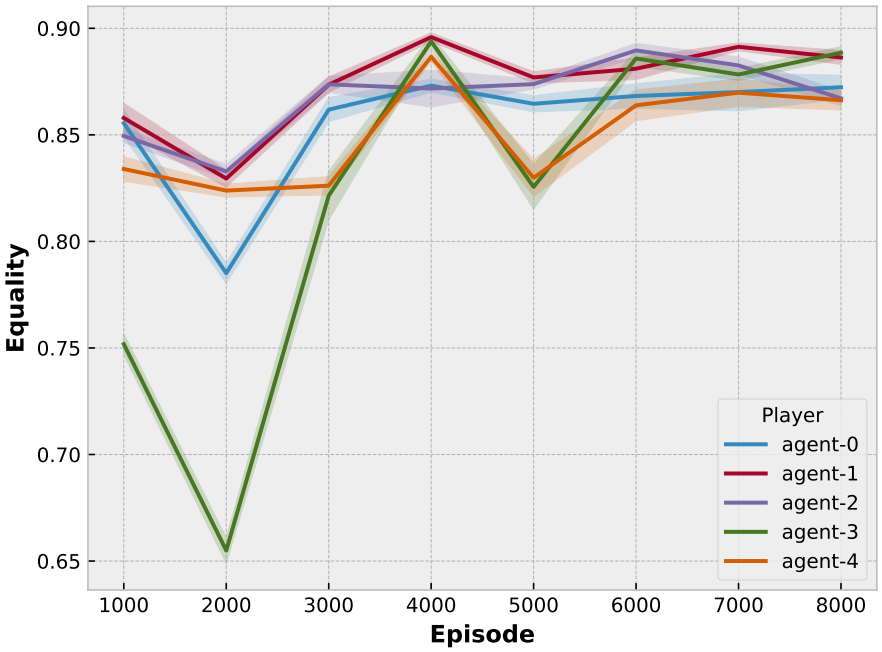


Figure 5: Evolution of Equality per agent over several episodes (averaged on 4 runs).

When analyzing these figures, it is clear that Efficiency (Fig. 3) is identical to the plot of Shapley values (Fig. 9 of the article). That is a logical result since the efficiency per agent (Eq. 1) computes the same result than Shapley values: the average contribution (reward) per agent. Equality (Fig. 5) also follows the same trend as Shapley values. This visual correlation supports our claim that Shapley values is a relevant tool to assert the contribution of agents in RL cooperative settings.

Experimental Details

Here, we present some additional details about the setup of the experiments showcased in Section V of the article.

Parameter	Value
Learning Rate	0.01
Optimizer	Adam
Number of MLP units	128
Discount Factor	0.95
Batch Size	1024

Table 1: Hyperparameters used for every MADDPG and DDPG model on Predator-Prey scenario. These are the default parameters recommended by [1]. Other hyperparameters (e.g., the number of predators or their speed) may vary and their values are indicated in the experimental settings description (Subsections V.B.1 and V.C.1 of the article).

Parameter	Value
Learning Rate	0.0001
Optimizer	Adam
Number of MLP units	128
Discount Factor	0.99
Batch Size	30000

Table 2: Hyperparameters used for A3C models on Harvest scenario. These are the default parameters recommended by [4]. Other hyperparameters (e.g. the number of agents) may vary and their values are indicated in the experimental settings description (see Subsection V.D.1).

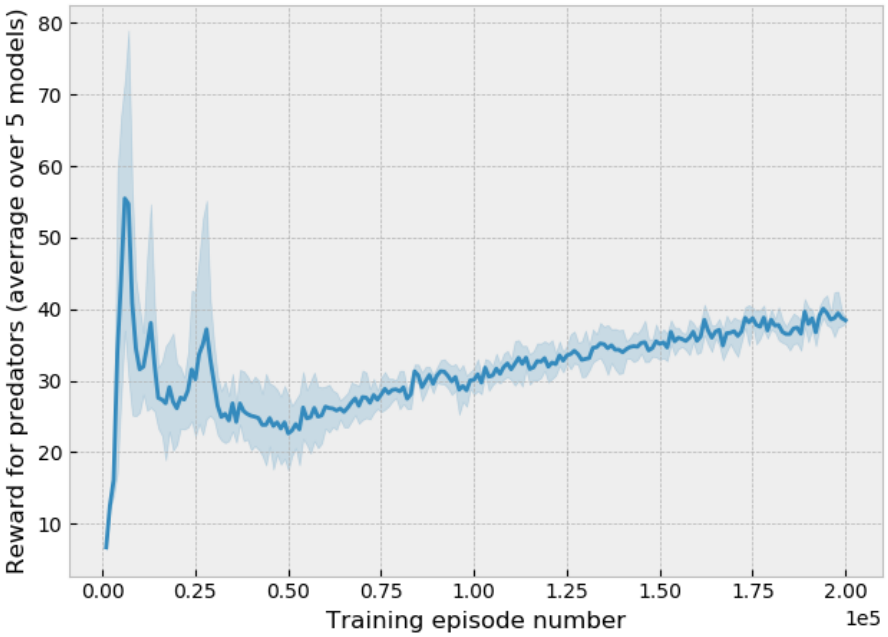


Figure 6: Learning curve of the Prey-Predator MADDPG models (average over 5 models) used in Experiment 1.

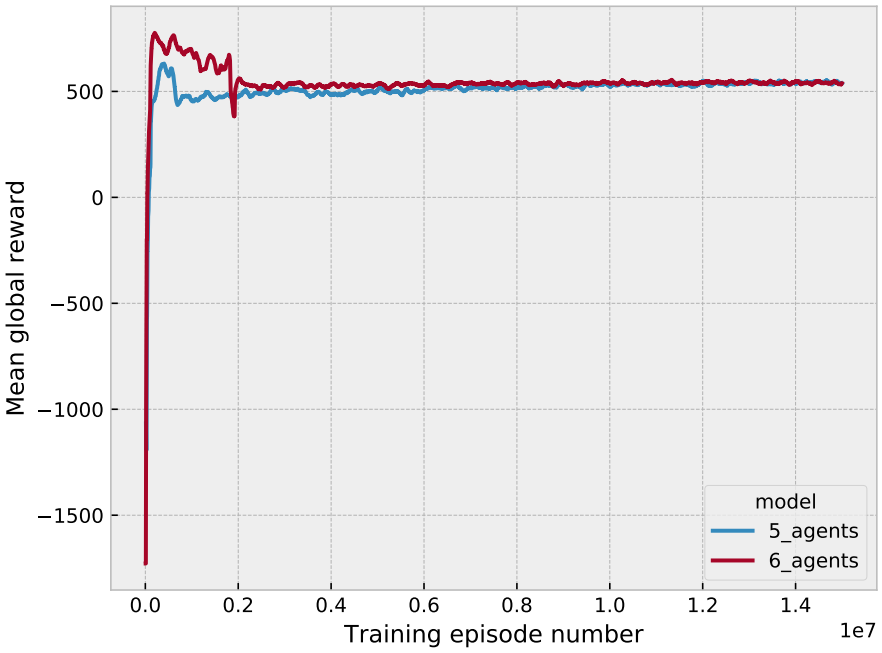


Figure 7: Learning curve of the Harvest A3C models (5 and 6 agents models used in Experiment 3). We can clearly observe that both models quickly converge to the same reward.

References

[1] Ryan Lowe et al. "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments". In: *Neural Information Processing Systems (NIPS)* (2017). URL: `https://arxiv.org/pdf/1706.02275.pdf`.

[2] Julien Perolat et al. "A multi-agent reinforcement learning model of common-pool resource appropriation". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 3646–3655.

[3] Peter Sunehag et al. "Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward". In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2018, pp. 2085–2087.

[4] Eugene Vinitsky et al. *An Open Source Implementation of Sequential Social Dilemma Games*. `https://github.com/eugenevinitsky/sequential_social_dilemma_games/issues/182`. GitHub repository. 2019.