Tunable Behaviours in Sequential Social Dilemmas using Multi-Objective Reinforcement Learning*

Extended Abstract

David O'Callaghan
School of Computer Science
National University of Ireland Galway
ocallaghan1@gmail.com

Patrick Mannion
School of Computer Science
National University of Ireland Galway
patrick.mannion@nuigalway.ie

ABSTRACT

In this study, we leverage multi-objective reinforcement learning to create tunable agents, i.e., agents that can adopt a range of different behaviours according to the designer's preferences, without the need for retraining. We apply this technique to sequential social dilemmas, settings where there is inherent tension between individual and collective rationality. Learning a single fixed policy in such settings leaves one at a significant disadvantage if the opponents' strategies change after learning is complete. In our work, we demonstrate empirically that the tunable agents framework allows easy adaption between cooperative and competitive behaviours in sequential social dilemmas without the need for retraining, allowing a single trained agent model to be adjusted to cater for a wide range of behaviours and opponent strategies.

KEYWORDS

Tunable agents; reinforcement learning; multi-objective decision making; multi-agent systems; social dilemmas

ACM Reference Format:
David O'Callaghan and Patrick Mannion. 2021. Tunable Behaviours in Sequential Social Dilemmas using Multi-Objective Reinforcement Learning: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 TUNABLE AGENTS

The standard approach to developing a reinforcement learning (RL) agent is to learn some fixed behaviour that will allow the agent to solve a sequential decision making problem. If, however, the developer wants the agent to behave differently, the agent normally has to be partially or completely retrained. To address this shortcoming, Källström and Heintz [2] introduced a framework to train agents whose behaviour can be tuned during run-time using multi-objective reinforcement learning (MORL) methods. In this framework, each set of objective preferences (scalarisation weights) corresponds to different combinations of desired agent behaviours, and the agent is trained with different weight vectors to learn different behaviours simultaneously. After the agent is trained, the weights can be adjusted on the fly to dynamically change the agent's behaviour, without the need for retraining. In this study we build on this framework, extending it to more complex environments with larger state-spaces and multiple learning agents.

In particular, we are interested in studying the suitability of the tunable agents framework to learn adaptive agent behaviours in sequential social dilemmas (SSDs) [3], settings where there is an inherent conflict between individual and collective welfare. In such settings, agents may choose to work together (cooperate) and share the resources available in the environment, or be selfish (compete) and attempt to maximise their own share of the resources, without considering the welfare of the other agents.

The training scheme employed by Källström and Heintz [1, 2] for training agents with tunable behaviours combines linear scalarisation of rewards [7] with the deep Q-network (DQN) algorithm [4]. At each time-step the agent samples a set of objective preferences (the weight vector) and updates the parameters of the neural network that is conditioned on both the current state of the environment and the weight vector using the scalarised reward as the target. After training, the tunable agent's behaviour can subsequently be adjusted by changing the weight vector.

Our aim in this study is to establish how effective the tunable agents framework first introduced by Källström and Heintz [2] is for developing agents that are capable of adjusting their degree of cooperation in SSDs. To this end, we conduct experiments in a modified version of the Wolfpack environment [3], an SSD where multiple predator agents aim to capture a prey. Our empirical results demonstrate that it is possible to train a single tunable agent that can easily adapt between cooperative and competitive behaviours in sequential social dilemmas without the need for retraining, catering for a wide range of possible behaviours and opponent strategies.

2 METHODOLOGY AND RESULTS

The environment used in our work is a 16×16 pixel-based gridworld (shown in Figure 1a). The predators are represented by blue pixels, the prey is represented by a red pixel and obstructions are represented by grey pixels. At each time-step, the predators and prey can move up, down, left, right or remain in their current location. As is the case in the original Wolfpack environment [3], the prey is captured if one of the predators moves to the same location as the prey. The capture is a team-capture if both predators are within a certain distance of the prey at the time of the capture. This distance is known as the capture-radius (shown in green in Figure 1a). Otherwise, the capture is a lone-capture. The Wolfpack environment was converted to a multi-objective stochastic game [6] by returning a reward vector to the predators, where one element corresponds to a team-capture and another corresponds to a lone-capture.

*An extended version of this paper is available [5].

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

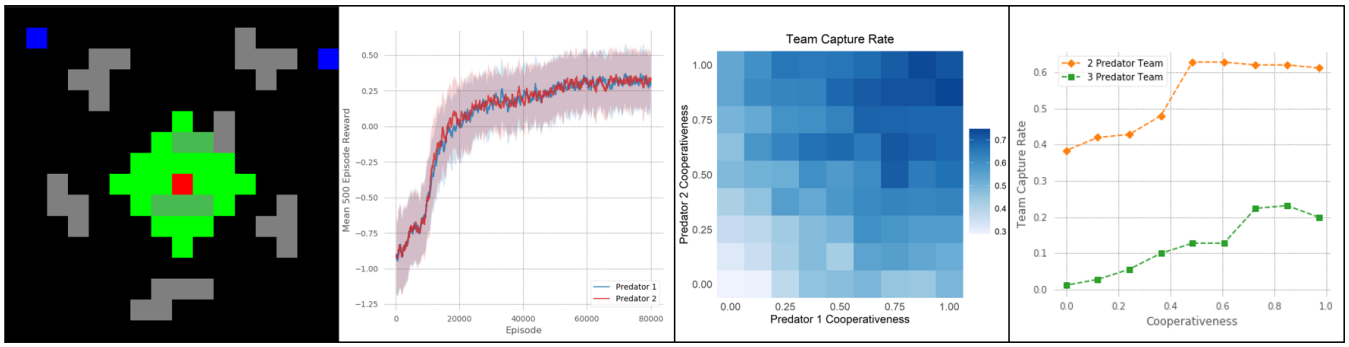

Figure 1: (*Left-to-right*) (a) Multi-objective Wolfpack environment, (b) training progress for tunable agents, (c) tuning performance for 2 predators with varied preferences, (d) tuning performance for 3 predators with matched preferences.

The training scheme used by Källström and Heintz [2] was adapted for this multi-agent environment by training a tunable DQN for each predator simultaneously and sampling objective preferences for each agent separately at the beginning of each episode. Each agent treats the other agents as part of the environment. Each predator agent sees themselves as a blue pixel and any other predators as green pixels so that each neural network can learn which pixel they represent [4]. The scalarised reward per episode during training is shown in Figure 1b.

After training, simulations in the Wolfpack environment with the trained tunable predators were viewed. The impact of tuning the objective preference weights on the behaviours of the predators was clear. When a high objective preference weighting was given to lone-captures, the predators would be competitive and move as quickly as possible towards the prey to capture it on their own. Conversely, setting a high weighting for team-captures resulted in the predators behaving cooperatively and approaching the prey together[1].

A quantitative analysis of the impact of tuning the predator agents' weight vectors was also completed. The team-capture rate over 250 episodes for each possible match-up of levels of cooperativeness (from an evenly spaced set of values) for the two predators was computed. The results from this experiment are displayed as a heatmap in Figure 1c. These results clearly demonstrate that the team-capture rate varies depending on the level of cooperativeness of each predator. The fact that the team-capture rate can be increased by increasing only one of the predators' levels of cooperativeness shows that the predator models don't learn any assumption of how the other predator behaves and that they generalise to behaviours of any level of cooperativeness. This is a direct result of sampling weight vectors for each predator independently during training. Note also that some of these levels of cooperativeness were not seen by the predators during training. Therefore, a secondary finding of this experiment was that the tunable agents have generalised to unseen objective preference weights.

As an additional experiment, a third predator was added into the Wolfpack environment to test how well the models can generalise to unseen states. During training, the images of the grid that either of the networks see contain one red pixel (for the prey), one blue pixel (for the predator that the network is being trained for) and only one green pixel (for the other predator). However, adding a third predator during simulation means that an image would contain two green pixels, meaning that all states in this 3-predator experiment would not have been experienced by the models during training. The plot in Figure 1d, however, shows that the models can generalise to behaving cooperatively or competitively when there is a third predator present. In this experiment, all predators were given the same weight vectors in a given episode and team-captures were monitored for both 2-predator teams and 3-predator teams. Both metrics are positively correlated with increasing levels of cooperativeness. However, 3-predator team-captures are less frequent than 2-predator team-captures. This is because the agents only experienced 2-predator team-captures during training, and there is no additional reward for a 3-predator team-capture.

Please see the extended version of this paper [5] for more detailed experimental results, including comparisons between tunable predator agents and predator agents trained with fixed objective preferences. The source code to reproduce our experiments may be downloaded from https://github.com/docallaghan/tunable-agents.

The contributions of this work open the door for this method of training agents with tunable behaviours to be applied to a huge array of different problems. This framework would be beneficial for any RL problem where there is some degree of uncertainty over the desired type of agent behaviour. If an agent with fixed objective preferences was trained and it was then seen that the behaviour needed to be changed slightly, the agent would need to be retrained with new objective preferences. However, using the method for training tunable agents that is the focus of this study, the objective preferences could simply be fine-tuned after training. In future work, non-linear scalarisation functions could potentially allow tunable agents' preferences over behaviours to be represented in a manner that fits better with how utility is derived in real-world multi-objective decision making problems (e.g. utility functions are non-linear in situations where a minimum value must be achieved on each objective).

[1]These two types of behaviours can be seen in the video recordings at https://www.youtube.com/playlist?list=PLuIjfbXklqXbNY1tXl7gtEWj5J5pCH_Ub

REFERENCES

[1] Johan Källström and Fredrik Heintz. 2019. Multi-Agent Multi-Objective Deep Reinforcement Learning for Efficient and Effective Pilot Training. In *FT2019. Proceedings of the 10th Aerospace Technology Congress, October 8-9, 2019, Stockholm, Sweden*. 101–111.

[2] Johan Källström and Fredrik Heintz. 2019. Tunable dynamics in agent-based simulation using multi-objective reinforcement learning. In *Adaptive and Learning Agents Workshop (ALA-19) at AAMAS, Montreal, Canada, May 13-14, 2019*. 1–7.

[3] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (São Paulo, Brazil) *(AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 464–473.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[5] David O'Callaghan and Patrick Mannion. 2021. Exploring the Impact of Tunable Agents in Sequential Social Dilemmas. *arXiv preprint arXiv:2101.11967* (2021). https://arxiv.org/abs/2101.11967

[6] Roxana Rădulescu, Patrick Mannion, Diederik M Roijers, and Ann Nowé. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34, 1 (2020), 10.

[7] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48 (2013), 67–113.