Learning Complex Teamwork Tasks Using a Given Sub-task Decomposition

Elliot Fosong
University of Edinburgh

Edinburgh, United Kingdom

e.fosong@ed.ac.uk

Ignacio Carlucho
Heriot-Watt University

Edinburgh, United Kingdom

ignacio.carlucho@hw.ac.uk

Arrasy Rahman
University of Texas at Austin

Austin, TX, USA

arrasy@cs.utexas.edu

Stefano V. Albrecht
University of Edinburgh

Edinburgh, United Kingdom

s.albrecht@ed.ac.uk

ABSTRACT

Training a team to complete a complex task via multi-agent reinforcement learning (MARL) can be difficult due to challenges such as policy search in a large joint policy space, and non-stationarity caused by mutually adapting agents. To facilitate efficient learning of complex multi-agent tasks, we propose an approach which uses an expert-provided decomposition of a task into simpler multi-agent sub-tasks. In each sub-task, a subset of the entire team is trained to acquire sub-task-specific policies. The sub-teams are then merged and transferred to the target task, where their policies are collectively fine-tuned to solve the more complex target task. We show empirically that such approaches can greatly reduce the number of timesteps required to solve a complex target task relative to training from-scratch. However, we also identify and investigate two problems with naive implementations of approaches based on sub-task decomposition, and propose a simple and scalable method to address these problems which augments existing actor-critic algorithms. We demonstrate the empirical benefits of our proposed method, enabling sub-task decomposition approaches to be deployed in diverse multi-agent tasks.

KEYWORDS

Multi-agent reinforcement learning; Ad hoc teamwork; Multi-agent transfer learning; Deep reinforcement learning

1 INTRODUCTION

In cooperative multi-agent reinforcement learning (MARL) [1], the goal is to have a team of autonomous agents learn to complete a task, by having the team gather and learn from experiences in that task. Although MARL techniques have been used successfully to solve a range of cooperative team-based tasks, there are still challenges in complex scenarios. These challenges include multi-agent credit assignment [5], non-stationarity due to simultaneously adapting agents [19], difficulty searching over a large joint action space, and equilibrium selection problems [6, 36]. These problems typically worsen when the number of agents increases, or when complex coordination is required.

We propose addressing these problems and solving complex multi-agent tasks by using a curriculum of sub-tasks. We focus on complex cooperative tasks that may be decomposed into sub-tasks, where each sub-task could be solved by a subset of the agents. We start by training sub-teams of agents on their respective sub-tasks before fine-tuning the full set of agents on the target task. In doing so, we induce a curriculum for the agents, allowing them to bypass the initial stages of random search by leveraging skills acquired during the first phase of training. Furthermore, by training initially in simpler tasks with fewer agents, we reduce the problems caused by non-stationary and multi-agent credit assignment. While we assume this sub-task decomposition is given, many tasks have a natural decomposition into sub-tasks that could use our proposed methodology.

For example, consider training five agents to play a 5-a-side football target task by breaking the problem up into two sub-tasks: attack drills with two attackers, and defence drills with two defenders and one goalkeeper (Figure 1). The attackers learn skills including "shooting on target" and "avoiding being tackled" which are useful in the full 5-a-side football game. Likewise, defenders learn skills useful in 5-a-side football such as "blocking" and "tackling". When the attackers and defenders are recombined, extra fine-tuning is required in the full 5-a-side football game, for example, to teach defenders that they ought to pass to their attacker teammates.

Code and experimental data available at https://github.com/uoe-agents/MEDoE
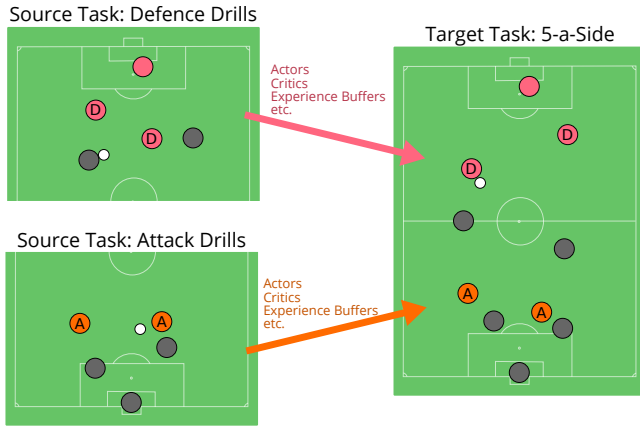
Figure 1: Sub-task curriculum for 5-a-side football. We control the red team, and play against the grey team. We train defender agents in defensive drills (top-left) and attacker agents in attack drills (bottom-left). As represented by the arrows, we transfer these agents to the 5-a-side target task. We then fine-tune our combined team in the target task.

We conjecture that MARL approaches can learn complex target tasks more efficiently by first training policies in each sub-task, followed by combining the sub-task policies and fine-tuning them in the target task. We explore the feasibility of this approach in experiments with human-designed sub-task curricula using standard MARL techniques building on top of the proximal policy optimisation (PPO) algorithm [25]. We find that using such an approach, we can learn to solve multi-agent tasks which are difficult for current state-of-the-art MARL algorithms, suggesting an interesting avenue for future MARL research and practitioners.

However, we also identify two issues with the naive application of standard MARL techniques to fine-tuning: (1) miscoordinated exploration leading to convergence to sub-optimal equilibria; and (2) agents forgetting useful skills they obtained during sub-task training. To address these issues, we further propose an approach called *Modulating Exploration and Training via Domain of Expertise (MEDoE)* which automatically infers based on sub-task experience replay buffers the circumstances in which an agent has expertise in the complex target task; and then uses the expertise predictions to modulate hyperparameters in the fine-tuning process. MEDoE is a flexible approach that can extend any decentralised execution actor-critic method, and can be used with any number of agents. Furthermore, MEDoE can also be used in situations where the size of the team in the complex target task may not be known during training in the simple sub-tasks, allowing for flexibility with respect to the target team composition, and re-use of pre-trained agents in fine-tuning for different downstream target tasks.

Continuing our football example, MEDoE might predict that in a defensive scenario in 5-a-side football (Figure 1, right) the defender agents have expertise — i.e., their policies are already near-optimal. Assuming these defenders do indeed have expertise, MEDoE reduces the likelihood the defenders take exploratory actions, reducing variance caused by exploration, thereby stabilising training of

other agent policies. Additionally, MEDoE reduces the rate at which the defenders forget their existing skills by increasing the coefficient of penalty terms in the policy loss for deviating from their final sub-task policies, and decreasing the entropy bonus coefficient.

Our experiments show that sub-task decomposition methods can solve complex teamwork tasks in many fewer training timesteps than baseline methods which train from scratch in the target task. In some environments, naive fine-tuning approaches are sufficient to outperform from-scratch baselines, but in others our proposed method MEDoE is responsible for the improved performance of using sub-task decompositions.

2 PROBLEM FORMULATION

In this section, we define the framework for our sub-task decomposition approach to accelerating MARL.

2.1 Sub-task Decomposition

We describe our sub-task decomposition approach in terms of *sub-task curricula (STCs)*, which model how a sub-task decomposition can be used in training. In this work, we consider cases where we are provided with a sub-task curriculum, and distinguish our work from those which seek to learn the curriculum. A sub-task curriculum C is a tuple, $\langle \mathcal{T}, \{C_{\text{sub.}}\}, A, \mathbb{L} \rangle$. The task $\mathcal{T}$ is a Dec-POMDP[17],

$$\mathcal{M} = \langle I, \mathcal{S}, \{\mathcal{A}_i\}_{i \in I}, T, \mu, \{\Omega_i\}_{i \in I}, O, R, \gamma \rangle, \quad (1)$$

where I is the set of agents (the team); $\mathcal{S}$ is the state space; $\mathcal{A}_i$ is the action space for agent i; $T(s_{t+1}|s_t, a_t)$ is the state transition probability density function; $\mu(s_0)$ is the initial state distribution; Ω_i is the observation space for agent i; $O(o_t|s_t, a_{t-1})$ is the observation probability density function; $R : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the team reward function; and γ is the discount factor. The objective within a Dec-POMDP is to find a joint policy π which maximises the expected discounted return $G = \mathbb{E} \left[\sum_t \gamma^t R(s_t, a_t) \right]$.

$\{C_{\text{sub.}}\}$ is a set of sub-task curricula, where each sub-task curriculum is defined recursively. The base-case for recursion, $\{C_{\text{sub.}}\} = \emptyset$, corresponds to training from scratch on task $\mathcal{T}$. We also refer to sub-tasks as *source tasks*, using transfer learning terminology [7].

A is a function which maps from sub-tasks $\{C_{\text{sub.}}\}$ to a set of agents I for task $\mathcal{T}$. Each agent has associated data — in this paper we use actor-critic methods, so each agent has an associated policy π, value function V, and we also include an experience replay buffer for each agent. For conciseness we also use A to represent the set of agents which initialise the current task in the curriculum.

$\mathbb{L}$ is a learning algorithm applied in each task of the sub-task curriculum, which given the agents A and Dec-POMDP $\mathcal{T}$ learns policies for task $\mathcal{T}$. For example, in this paper $\mathbb{L}$ is the IPPO algorithm with particular hyperparameters and stopping conditions.

Though we present a general framework for modelling sub-task curricula, in this work, we focus on curricula with tree-depth of one. As a concrete example, the 5-a-side football task from Figure 1 the sub-task curriculum would be:

- $C = \langle \mathcal{T}_{5v5}, \{C_{\text{def.}}, A, C_{\text{att.}}\}, \mathbb{L} \rangle$

- $C_{\text{def}} = \langle \mathcal{T}_{\text{def.}}, \emptyset, A, \mathbb{L} \rangle, C_{\text{att.}} = \langle \mathcal{T}_{\text{att.}}, \emptyset, A, \mathbb{L} \rangle$

2.2 Objective

Our objective is to accelerate the rate at which MARL learns to solve a complex target task. Here we define our objective in terms of minimising the total number of timesteps taken to reach a desired level of performance in the target task.

Let $N(C)$ be the number of training timesteps in the current task of the curriculum C, and $N^{\text{tot.}}(C)$ be the *total* number of training timesteps used in the entire sub-task curriculum C_i, i.e.,

$$N^{\text{tot.}}(C) = N(C) + \sum_{C_s \in \{C_{\text{sub.}}\}} N^{\text{tot.}}(C_s)$$

Given a target task $\mathcal{T}_{\text{target}}$ and target performance G_{target} our overall objective is to find a sub-task curriculum $C^*_{\text{target}} = \langle \mathcal{T}_{\text{target}}, \cdot, \cdot, \cdot \rangle$ which minimises $N(C^*_{\text{target}})$ subject to the constraint that the produced agents attain expected returns $G \geq G_{\text{target}}$.

This optimisation problem has many free variables including the learning algorithm used at each point in the curriculum; the stopping conditions at each point in the curriculum; the sub-tasks used at each point in the curriculum; and the agents transferred at each point in the curriculum. In this work, we first investigate the feasibility of the approach based on standard MARL components without attempting to optimise the sub-task curriculum. We then propose an approach to improve performance by modifying the learning algorithm $\mathbb{L}_{\text{target}}$ used in the target task.

3 SUB-TASK CURRICULA WITH STANDARD MARL APPROACHES

In this section, we investigate applying standard MARL techniques for fine-tuning in the target tasks of sub-task curricula. We find that while this naive approach can sometimes work, we identify and analyse two problems which can arise. In Section 4 we present a method which addresses these problems and greatly improves the performance of sub-task curriculum approaches in some tasks.

3.1 Environments

To test the sub-task curriculum approach, we consider three environments with clear task decompositions: *Chainball*, a simple but difficult to solve environment we introduce to provide insight; *Overcooked* [24, 33], a gridworld environment common in MARL research; and *VMAS Football* [2], a complex 2D physics-based football simulation. These tasks are fully-observable, and we ensure that observation dimensions are consistent between tasks within the curriculum by zero-padding observations where necessary. Further details of each environment can be found in **??**.

Chainball. We introduce the *Chainball* environment as test-bed for the sub-task curriculum approach, mimicking the compositional properties of our football motivating example, while being cheap to evaluate. Chainball is a difficult task for MARL due to the sparse reward and high degree of coordination required. Chainball (Figure 2a) is an episodic 4-player game with $|\mathcal{S}| = 11$ discrete states. The goal of Chainball is to reach the rightmost end of the chain ("goal scored"), where the team will receive a reward of +1. However, if the leftmost end of the chain is reached ("goal conceded"), the team receives a reward of -1. Upon scoring, the state is reset to

the middle state (s_6), and the episode terminates after 90 timesteps. The transition probability is defined by a matrix for each state — at timestep t, each of four agents chooses an action $a_{t,i} \in \{1, 2, 3, 4\}$, and the probability of moving right is given by the corresponding matrix entry for that state.

Chainball has two 2-player source tasks, Chainball-Att and Chainball-Def, to emulate attack and defence drills respectively. Each source task consists of 11 states, but we make $s < 7$ in Chainball-Att and $s > 5$ in Chainball-Def terminal states. Source task episodes also terminate if a goal is scored or conceded.
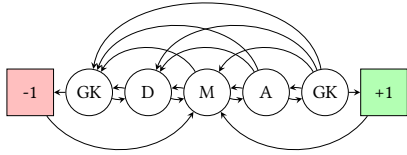
Overcooked. We use the *Overcooked* [24] environment, which requires multi-step coordination. The goal of Overcooked is to complete a recipe by moving and processing foods in a grid world. Figure 2b shows the configuration of our Overcooked sub-task curriculum. In the target task, agents must coordinate to pass and chop the tomato on the chopping board (1,2), put the chopped tomato on a plate (3), and pass the plate with the chopped tomato back to serve at the starred counter (4,5). The skills to complete steps 2 and 3 can be learned in the "Right" task; and skills to complete step 5 can be learned in the "Left" task. Steps 1 and 4 require learning new behaviour in the target task. The team is rewarded for completing each step in the recipe, except steps 1 and 4 in the target task.

VMAS Football. The *VMAS Football* environment [2] is a 2D physics-based version of football. VMAS Football has a discrete action space (control inputs in four cardinal directions). Our sub-task curriculum in VMAS football closely follows that shown in Figure 1, except we do not use goalkeeper agents. We use a sparse reward of +1 for scoring and −1 for conceding. Episodes terminate when either team scores, or after 1024 timesteps. The opponent team uses the heuristic policy provided by the VMAS environment.
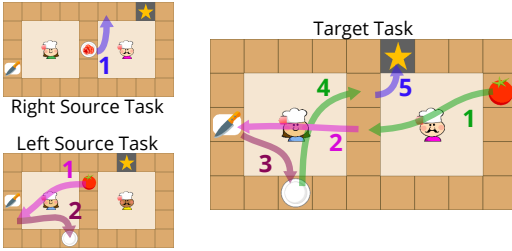
3.2 Protocol and Baselines

In our initial experiments we consider two settings: 1. training from scratch in the target task, and 2. training using a simple sub-task curriculum approach.
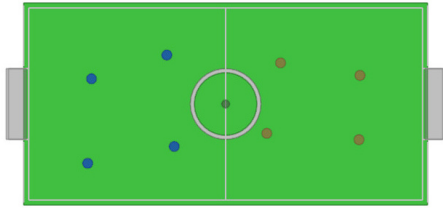
(1) For our "From-Scratch" baselines, we consider the state-of-the-art MARL algorithms multi-agent PPO (**MAPPO**) [39] and **QMIX** [23]. We also test the independent PPO (**IPPO**) [20] algorithm in a from-scratch setting, as an ablation of our sub-task curriculum method which uses IPPO. We run each of these algorithms for 16 seeds directly on the target task. Due to computational constraints, in VMAS Football we use 8 seeds.

(2) To investigate the utility of a sub-task curriculum (STC), we use a simple approach extending IPPO, which we label **Naive STC**. For each source task, we generate four seeds of skilled subteams by training agents using the standard IPPO algorithm (without parameter sharing) until convergence. In Chainball, we use tabular actors and critics; whereas in Overcooked and VMAS Football we use neural networks. We initialise the target task team networks by cloning final actors and critics from the source tasks. This gives 16 final teams, formed by the 4^2 pairings of source task teams. Given these initialisations, Naive STC then fine-tunes in the target task using the standard IPPO algorithm, and we average the results over these 16 runs. For

(a) 5-state Chainball Environment. Here states are labelled with GK (goalkeeper), D (defence), M (midfield), and A (attack).

(b) Overcooked sub-task curriculum. Arrows represent cooking steps within each task.

(c) VMAS Football target task. The sub-task curriculum used is similar to that shown in Figure 1.

Figure 2: Environments used in our experiments.

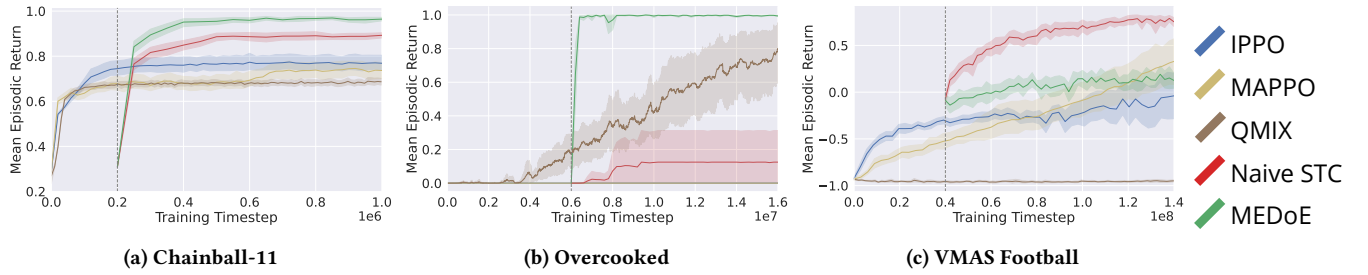| State	1	2	3	4	5	6	7	8	9	10	11
Sub-optimal	90	63	54	38	2	0	2	6	4	23	54
PD NE	7	3	38	27	0	-	0	0	50	91	58
PD ϵ-NE	23	3	38	56	0	-	0	67	50	91	81

Table 1: Chainball equilibrium analysis for Naive STC, showing for each state the percentage of runs where the agents' most likely joint action is sub-optimal; and the percentage of those sub-optimal joint-actions which are also Pareto-Dominated Nash Equilibria (PD NE) or ϵ-NE with $\epsilon = 0.05$.

continue to investigate these approaches by asking *why* they sometimes fail to perform as well as hoped. Via analysis of our chosen test domains, we conjecture two pitfalls: coordination difficulties arising from miscoordinated exploration (Section 3.4); and issues caused when agents forget useful behaviours they learned during source task training (Section 3.5). In Section 4, we propose an algorithm designed to address these issues, and show that it leads to an improvement in performance in Chainball and Overcooked.

3.4 Miscoordinated Exploration

The first problem with naive sub-task curricula we identify is most evident in the Chainball environment — miscoordinated exploration, which can lead to convergence upon *Pareto-dominated Nash equilibria* which is a well-known cause of coordination failure in multi-agent learning systems [6, 10]. Pareto-dominated Nash Equilibria in cooperative games are Nash equilibria which are Pareto-dominated by some other Nash equilibrium. That is, a joint strategy in which no agent has the incentive to *unilaterally* deviate from, but where there exists some other Nash equilibrium with a higher return to *all* players. It is possible that learning agents can get "stuck" in Pareto-dominated Nash equilibria if the risks of occasional deviation from the equilibrium strategy are larger in the Pareto-dominated Nash equilibrium than the Pareto-optimal Nash equilibrium. A classical example of this is the Stag-Hunt game, where independent learning agents – particularly those which take exploratory actions which cause them to sometimes deviate from the optimal (Stag, Stag) equilibrium – often converge upon the sub-optimal (Hare, Hare) equilibrium in practice.

Recall that at each timestep, Chainball has a matrix-game structure where agents jointly choose an action which determines the probability with which the players move towards the opponent's goal. This allows us to perform an equilibrium analysis to demonstrate this issue by examining the policies obtained by fine-tuning agents in the target task using Naive STC, as shown in Figure 3a. Consider the most likely joint action given by the policies at the end of fine-tuning. For each state, we count the fraction of runs (over our 48 Chainball runs) where the most likely joint action is not the optimal action in that state. We report results in Table 1. Focussing on states 10 and 11 (as the lower states are less frequently visited), we see that convergence to Pareto-dominated Nash equilibria is particularly common in Chainball, and leads to degraded performance of sub-task curriculum methods.

Chainball we use 3 seeds per fine-tuning stage, giving a total of 48 runs. Plots in this section also show results for our proposed approach named **MEDoE**, which we discuss in Section 4.

We describe our hyperparameter tuning protocol and report chosen hyperparameters in **??**.

3.3 Results

Our experiments in Figure 3 show mixed results for the Naive STC approach relative to standard MARL baselines. In Chainball, we observe that the use of a sub-task curriculum allows the agent to receive a higher converged return than all the from-scratch baselines. Nevertheless, Naive STC falls short of achieving the maximum attainable return of 1.0. In Overcooked, Naive STC outperforms IPPO and MAPPO, which do not learn at all in any of their 16 runs. However, it does not perform as well as QMIX from-scratch. In VMAS Football, using a sub-task curriculum enables finding a solution to the task in many fewer timesteps than from-scratch methods.

Contrary to the expectation that the use of a sub-task curriculum can make learning a complex target task more efficient, our results show that this is not always the case. This may explain why, despite the simplicity, naive sub-task curriculum approaches have not been widely used or investigated in the MARL literature. We

(a) Chainball-11

(b) Overcooked

(c) VMAS Football

Figure 3: Target task training returns for each environment. Mean episodic returns (100 episodes), averaged over 16 runs (16 different team combinations). The "from-scratch" baselines are averaged over 16 runs (16 seeds; 8 seeds for QMIX on VMAS Football). Shaded area shows the 95% confidence interval of the mean over the runs. Naive STC and MEDoE are shifted on the training step axis to account for the total training steps required across all the sub-tasks, shown by the dashed line.

(a) Left Half

(b) Right Half
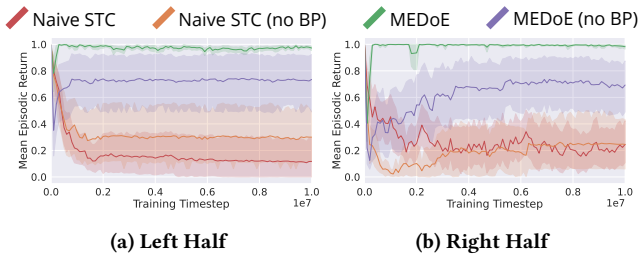
Figure 4: Forgetting curve in Overcooked.

3.5 Forgetting

We identify a problem prominent in Overcooked — agents forgetting useful behaviours. Consider the Overcooked target task (initial state shown in Figure 2b). The "right" agent needs to learn step 1. However, until it does so, neither agent receives a reward signal. As we use entropy regularisation, this means that both agents' policies will gradually reset towards uniform policies until eventually steps 1 and 2 are completed. We hypothesise that this gradual resetting causes agents to gradually forget skills relevant to steps 2, 3, and 5.

To investigate this hypothesis, we perform an experiment in which we take agents trained in source tasks and fine-tune them in the target task. As we fine-tune the agents, we periodically re-evaluate the agents *in their respective source tasks* to test whether they have retained the skills learned in their source task. The results shown in Figure 4 show evidence of forgetting — as agents are fine-tuned on the target task, their sub-task performance drops, corresponding to a loss of skill on the sub-task.

4 MODULATING EXPLORATION AND TRAINING VIA DOMAIN OF EXPERTISE

In this section, we introduce a novel approach, *Modulating Exploration and Training via Domain of Expertise (MEDoE)*, designed to facilitate efficient learning in the fine-tuning stage of a sub-task curriculum. We focus on addressing the identified problems of miscoordinated exploration (Section 3.4) and forgetting (Section 3.5). MEDoE comprises two key components: a module which predicts based on source task information whether an agent's policy is likely to be optimal in the current state; and a module which takes this

prediction of expertise to modulate exploration and training hyperparameters. The key intuitions motivating MEDoE are that i) if an agent's policy is already near-optimal, it should explore less aggressively; and ii) if an agent's policy is near-optimal then the policy update should be regularised to stay close to this policy in order to prevent forgetting. In the following subsections, we discuss each of the two components of MEDoE in turn. We then carry out experiments to show that MEDoE can indeed address these issues.

4.1 Domain of Expertise Classification

The first component of MEDoE is the domain of expertise (DoE) classifier, which for each agent makes a prediction about whether that agent's policy is already close to the optimal decentralised policy for the target task. MEDoE makes these predictions based on information from the source tasks, using the heuristic that if a target-task state is similar to states observed in an agent's source task, then that agent's policy is more likely to be near-optimal.

We formalise the notion of DoE as follows: Let π_i be agent i's current policy. Let $\Pi^{*,T}$ be the set of optimal decentralised policies of the target task $\mathcal{T}_T$. We consider a target task observation $o_i \in \Omega_i^T$ to be in the *domain of expertise*, $\mathcal{E}_i^T$, of agent i if and only if

$$\exists \pi^{*,T} \in \Pi^{*,T}, \ D_{KL}(\pi_i(\cdot|o_i) \ \| \ \pi_i^{*,T}(\cdot|o_i)) < \tau, \qquad (2)$$

where τ is a similarity threshold, and D_{KL} is the KL divergence.

It is impossible to find an algorithm which improves generalisation performance on average across an unconstrained set of source and target tasks [37]. We therefore only expect MEDoE to work in cases where, at any given time, typically at least one agent's source task policy is optimal in the target task, i.e. $\bigcup_{i \in I^T} \mathcal{E}_i^T \approx \mathcal{S}^T$.

Knowing the ground-truth DoE set requires knowing the optimal policy, which is not practical. Instead in MEDoE we use experience buffers from each source tasks to infer a *DoE Classifier*, $D_i : \Omega_i^T \mapsto [0, 1]$ for each agent i. The ideal DoE classifier outputs

$$D_i^*(o_i) = \begin{cases} 1 & \text{if } o_i \in \mathcal{E}_i^T, \\ 0 & \text{if } o_i \notin \mathcal{E}_i^T. \end{cases} \qquad (3)$$

In practice, we formulate the DoE classifier training as a binary classification problem where positive examples are taken from agent i's source task experience buffer, and negative examples are taken from

the experience buffers of all agents trained in a different source task to i. We train a multi-layer perceptron (MLP) which learns to identify features of observations which differ between different source tasks, using the heuristic that if the current observation in the target task has features present in a particular source task, then it's likely the agents trained in that source task have expertise. For example, a classifier trained to distinguish between sample observations from attack drills and defence drills in football might identify the position of the ball on the pitch as a feature of interest. Defence drills would be identified by the ball being dear our team's goal, which likely corresponds with the DoE for defenders.

Concretely, let C_i represent the source task agent i was trained in. Each agent has an associated source task experience buffer $E_i = \{o_i^e : e \in 1 \ldots E_{\max}\}$. For each agent $i \in I^T$ we form a dataset $\mathcal{D}_i = \{\langle o_j^e, \mathbb{1}[C_i = C_j] \rangle : j \in I^T\}$. We then train an MLP classifier which uses a sigmoid final layer to output a probability $p \in (0, 1)$, with the objective of minimising binary cross entropy loss on $\mathcal{D}_i$.

4.2 Exploration Modulation

The second module of MEDoE is the modulation of the exploration and training process during target task fine-tuning, informed by the DoE classifier. In this paper, we focus on our variant of MEDoE based on PPO [25], which modulates three quantities:

1) to address forgetting (Section 3.5), MEDoE modulates a) the policy entropy regularisation coefficient, α and b) the policy behaviour prior coefficient, κ;

2) to address miscoordinated exploration (Section 3.4), MEDoE modulates the policy softmax action selection temperature, T.

In the following section, we discuss the intuition behind the modulation of these quantities, and describe MEDoE. We provide pseudocode for our algorithm in **??**.

Experts should retain relevant skills. and non-experts should forget irrelevant skills. During the source task, agents learn skills which are relevant to the completion of the target task, but also skills which might be irrelevant. Such irrelevant skills can arise from differences in skill requirements between source and target tasks, or from extrapolation. Ideally, agents should quickly forget irrelevant behaviours. However, at the same time, they must retain useful skills, which may be difficult in settings which require complex coordination or with sparse rewards, since forgetting can occur during extended low-reward periods.

To control the rate of forgetting skills, MEDoE modulates two parameters. Firstly, we use entropy-regularised policies, and encourage non-experts to forget irrelevant skills by increasing non-experts' entropy regularisation coefficient, setting $\alpha_i = \alpha_{\text{base}} \times \beta_\alpha^{(1-\hat{D}_i(o_i))}$, where α_{base} is the base entropy coefficient, and $\beta_\alpha \geq 1$ is the entropy boost coefficient introduced by MEDoE. A high entropy regularisation coefficient increases the rate at which a policy relaxes towards a uniform distribution, intuitively corresponding to an increased rate of forgetting skills.

Secondly, we use fixed behaviour priors [30] to encourage experts to retain useful skills. This entails using KL-regularised policies (see

Equation (6)), where we aim to minimise the KL divergence between the agent's current policy $\pi_i(a_i|o_i; \theta_i)$, and its frozen source task policy $\pi_i(a_i|o_i; \bar{\theta}_i^{\text{BP}})$, thereby encouraging the agent to stay close to its source task behaviour. We boost the KL regularisation coefficient for experts, setting $\kappa_i = \kappa_{\text{base}} \times \beta_\kappa^{\hat{D}_i(o_i)}$, where $\beta_\kappa \geq 1$.

Non-experts should explore. and experts should be predictable to other agents by exploiting existing skills. By definition, non-expert agents need to learn new behaviours. To do so they must explore. Exploration in multi-agent systems can have negative effects on learning, such as reducing training stability, and as discussed in Section 3.4, increasing the difficulty of selecting equilibria which require stable coordination. We therefore aim to restrict exploration to situations where it is necessary, i.e., when agents are non-experts. MEDoE takes a simple approach: modulate an agent's exploration parameter using that agent's DoE classifier. For the PPO-based MEDoE, the relevant exploration parameter is the stochastic action selection temperature T_i, which scales the policy softmax input:

$$a_i \sim \pi(\cdot|o_i; T_i = T_{\text{base}} \times \beta_T^{(1-\hat{D}_i(o_i))}), \ \forall i \in I, \tag{4}$$

where $\beta_T \geq 1$. During evaluation, we fix the action selection temperature to T_{base}. We therefore apply an importance sampling reweighting w_i (Equation (5)) during training:

$$w_i = \frac{\pi_i(a_i|o_i; \theta_i, T = T_{\text{base}})}{\pi_i(a_i|o_i; \theta_i, T = T_i)}. \tag{5}$$

Ultimately, we minimise the following policy and value losses for each agent i in the target team:

$$\begin{aligned} \mathcal{L}(\theta_i) =\ & w_i \text{PPOClip}(A_i, \pi_i(a_i|o_i; \theta_i), \epsilon) \\ & - \alpha_i H(\pi_i(\cdot|o_i; \theta_i)) \\ & + \kappa_i D_{KL}(\pi_i(\cdot|o_i; \theta_i) \parallel \pi_i(\cdot|o_i; \bar{\theta}_i^{\text{BP}})), \end{aligned} \tag{6}$$

$$\mathcal{L}(\psi_i) = w_i \|G_{t:t+n} - V_i(o_i; \psi_i)\|_2^2. \tag{7}$$

where $\text{PPOClip}(A, \pi, \epsilon)$ is the PPO policy ratio clipping function described by Schulman et al. [25] with clipping coefficient ϵ. The advantage function for agent i, $A_i = (G_{t:t+n} - V_i(o_{i,t}; \psi_i))$, and the n-step return for agent i, $G_{t:t+n} = \gamma^n V_i(o_{i,t+n}; \psi_i) + \sum_{i=0}^{n-1} \gamma^i r_{t+i}$, are defined in the usual way.

4.3 MEDoE Results

In this section, we present and discuss results from using our proposed approach, MEDoE. We investigate three questions: (1) whether MEDoE can improve the performance of using a sub-task curriculum; (2) whether behaviour priors alone are responsible for preventing forgetting; and (3) which of the MEDoE-modulated hyperparameters affect performance the most.

Does MEDoE improve performance? In Figure 3, we compare our approach, MEDoE to the naive sub-task curriculum approach described in Section 3. We find that for Chainball and Overcooked, MEDoE significantly outperforms the from-scratch baselines where the naive sub-task curriculum approach did not, solving the task in many fewer timesteps than all other baselines. This suggests that MEDoE can indeed address the problems identified in Section 3. However, we also find that MEDoE does not improve performance in VMAS Football relative to a naive sub-task curriculum, where
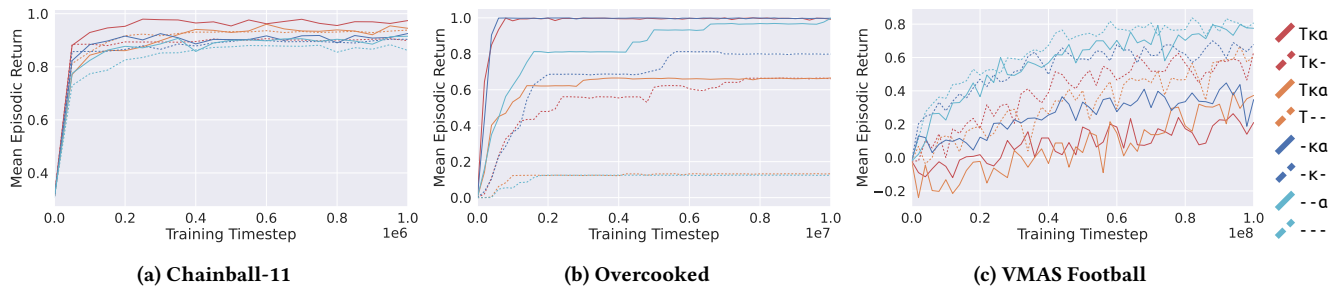
Figure 5: Ablations of MEDoE. Returns are averaged over 16 runs (4 runs in VMAS Football). Error bars are omitted to improve clarity. In the legend, labels show a letter if that coefficient is modulated, or a dash otherwise. E.g., "$T-\alpha$" modulates temperature coefficient (T) and entropy coefficient (α) but not KL coefficient (κ).

the naive approach already significantly outperforms from-scratch baselines. Our ablation study (Figure 5c) suggests that temperature modulation is particularly harmful to the performance of MEDoE in VMAS Football. We hypothesise that this is because the quasi-continuous nature of VMAS Football requires effective exploration over temporally extended sequences [18], whereas MEDoE's instantaneous increase in action selection temperature merely leads to poor performance (e.g., "dithering" causing an inability to consistently move in a straight line). Future work could therefore consider coupling a MEDoE-like approach with hierarchical RL to improve performance in (quasi-)continuous control settings.

Are behaviour priors sufficient to address forgetting? MEDoE adds a behaviour prior [30] term to the policy loss to prevent expert agents from forgetting useful skills. We investigate whether it is the behaviour priors alone which cause the improved performance due to preventing forgetting, or whether MEDoE provides benefits of its own. To do so, we present an additional baseline where we augment the naive sub-task curriculum approach with behaviour priors. Figure 3 shows that although the inclusion of behaviour priors alone does lead to increased performance of the naive sub-task curriculum approach, we find that MEDoE still outperforms this augmented baseline. This is particularly evident in Overcooked, which was the motivating case for controlling the rate of forgetting.

Secondly, results from the forgetting experiments (Figure 4) show that the use of behaviour priors does indeed slow the rate of forgetting, as expected. However, we also see that MEDoE enables agents to forget less than both from-scratch baselines and the behaviour prior augmented naive sub-task curriculum approach. Interestingly, with MEDoE we see a rapid drop in source task performance during the initial stages of target task fine-tuning, but MEDoE then rapidly recovers as the skills are recalled from the behaviour priors.

Which hyperparameters are most important to modulate? Our ablation experiments (Figure 5) provide insight into the role the different modulation parameters play in MEDoE. We find that the impact of the different parameters depends on the environment.

In Chainball, we see that each of the hyperparameters contributes somewhat towards MEDoE' improved performance. As anticipated, temperature modulation seems to be important to achieving the

highest returns in Chainball. In Overcooked we see that modulating the entropy and behaviour prior KL coefficients is crucial for MEDoE's performance, due to the importance of forgetting in Overcooked. Modulating the action selection temperature in Overcooked has little consistent effect, perhaps due to the lack of difficult equilibrium selection problems posed by Overcooked compared to Chainball. In VMAS Football, we see that temperature modulation is actually harmful to the performance of MEDoE. This provides evidence for our hypothesis (discussed earlier in this section) that "exploring" by boosting action selection temperature is a poor strategy in quasi-continuous settings like VMAS Football. However, we also find that modulating KL and entropy coefficients does not improve performance in VMAS Football. We believe this is particular to the policies learned in VMAS Football — we observe that a team can achieve high performance provided one agent learns to skilfully evade the opponent team to score, while the other agents simply defend in the rare cases it is necessary. This reduces the advantage to agents retaining their sub-task skills in VMAS Football.

5 RELATED WORK

To address challenges in multi-agent reinforcement learning, prior methods also investigated modulating training parameters. In WoLF-PHC [4] and extensions [3], each agent's policy learning rate is modulated according to the intuition that an agent's policy learning rate should be high when it is underperforming relative to its expectations, and low otherwise. MA2QL [27] focusses on a team learning setting. MA2QL tackles non-stationarity by allowing only one agent to learn at a time, rather than learning simultaneously.

Work by Vrancx et al. [31] considers transfer from simple tasks to complex target tasks in multi-agent systems. They train a classifier to distinguish between cases in which agents can learn individually, and cases in which they must learn to coordinate as part of a team. Though similar to our DoE classifier, one key difference is that our DoE classifier attempts to classify states in which further learning is not required, rather than states in which agents continue to learn without paying attention to other agents.

Wang et al. [35] also accelerates learning of complex multi-agent tasks using a curriculum. However, that work focusses on cases in which the number of agents is progressively increased throughout the curriculum. By contrast, our work considers cases in which the

task is decomposed based not upon the number of agents, but upon the different skills required by agents in the target task. Similarly, Tang et al. [28] consider scenarios in which agents join an unfamiliar team, and have to rapidly learn to adapt to coordinate with the new team to complete a known task. In contrast with our work, Tang et al. [28] vary only the number of agents between source and target tasks, while the underlying dynamics remain the same.

Taylor et al. [29] consider *parallel transfer learning*, which transfers experiences collected in parallel by separate agents into a target agent, similarly to federated reinforcement learning [22]. In contrast, our work does not directly transfer skills into individual agents, but instead attempts to accelerate the progress of the *team's* performance. Another transfer learning work by Yang et al. [38] provides a framework for transferring skills between agents within a task. In our work, one agent's knowledge and skill might not be useful to another agent — we instead focus on how each agent should use its own knowledge and skills in the new task.

Several single-agent RL methods use a multi-agent approach to curriculum learning [8, 15, 21]. Wang et al. [32] consider similar approaches in a MARL setting. However, these methods focus on the *curriculum design* problem (i.e., generating the series of tasks that form the curriculum) by treating it as a two-player game. We instead focus on the problem of accelerating learning at each stage in the curriculum, when the target task and decomposed sub-tasks are given as input by an expert. Our work can therefore be thought of as a *consumer* of sub-task decompositions. Future work could consider uniting MEDoE with methods that generate a sub-task decomposition. To our knowledge, there are few works which consider the challenge of automatic sub-task decomposition. Approaches using large language models [12] may be promising for learning to decompose tasks, using knowledge embedded in human text data.

The problem of effective fine-tuning on new tasks often appears in the continual learning literature, typically in single-agent settings. Nekoei et al. [16] propose using the card game Hanabi as a test-bed for continual learning in multi-agent settings. Liu et al. [13] engineer an approach to train humanoid agents to play 2-vs-2 football, which, like MEDoE, uses behaviour priors [30]. However, unlike MEDoE, their solution is complex and domain-specific.

Some works consider the assignment of agents into different roles, where agents assigned to the same role employ similar policies [34]. In our case, we assume roles do not have to be discovered, and are instead provided implicitly via the given sub-task curriculum. We then focus on efficient use of the sub-task curriculum, rather than training directly in the target task.

Finally, our work is inspired by Ad hoc teamwork (AHT): the problem of designing single agents capable of coordinating on the fly with previously unseen teammates [14, 26]. Fosong et al. [9] extended the AHT problem to the Few-Shot Teamwork (FST) problem wherein separate *teams* of agents must rapidly learn to complete a new task as part of a unified team. Fosong et al. provide two framings of the FST problem: an *ad hoc teamwork framing* and a *curriculum framing*. Though our work focuses on the latter framing by investigating methods to accelerate the MARL training process, MEDoE may also be applicable under the former framing.

6 CONCLUSION AND FUTURE WORK

In this work, we investigated the use of sub-task curriculum (STC) methods as an approach for accelerating learning of complex MARL problems. We found that given a decomposition of a complex teamwork task into simpler sub-task, straightforwardly applying the standard IPPO method to fine-tune agents trained in the sub-tasks can sometimes reduce the number of timesteps required to solve the complex task relative to state-of-the art MARL baselines which train from scratch on the complex task. However, we also showed that counter to expectations, the use of an STC does not necessarily improve performance over from-scratch MARL baselines, and does not necessarily converge to optimal policies. We identified and investigated two issues which hamper the performance of the naive STC approach: miscoordinated exploration and forgetting.

To address these issues, we presented Modulating Exploration and Training via Domain of Expertise (MEDoE), which modulates relevant hyperparameters of each agent during target task fine-tuning. Each agent's modulation is controlled by a domain of expertise classifier that provides information about whether the agent's policy is likely to be useful in the target task, given the current observation. We found that MEDoE converges to higher returns in fewer timesteps than the naive STC approach in two out of three domains.

MEDoE is a computationally cheap and scalable extension to existing centralised training decentralised execution (CTDE) actor-critic methods, with total computational cost increasing linearly with the number of agents in a parallelisable manner. Though our experiments extend the IPPO algorithm, other actor-critic methods can be modified to produce a MEDoE version, if desired. MEDoE does not require additional expert knowledge or engineering input beyond that required for a naive STC. Though it may not always improve performance in all tasks, MEDoE is a promising approach we recommend deploying whenever naive STC approaches fail.

Our findings open the door for future work on optimal use of sub-task curricula in MARL, and may be extended in multiple directions. Firstly, we assumed the target task and the sub-task curriculum was provided. In many cases, this may be a straightforward part of the training setup engineering, though sometimes it may be desirable to automatically propose sub-task decompositions based on task descriptions. This is an open challenge [11], though recent works employing large language models for this task (e.g. [12]) may be promising. Relatedly, future work in the vein of curriculum learning might focus on learning *optimal* sub-task curricula for a given task. Secondly, in our experiments we focus on one-step curricula, although the framework presented in Section 2.1 can represent curricula of arbitrary tree depth. Future work could investigate multi-step curricula, and whether any additional issues arise from their use. Finally, we present a simple scheme for learning a DoE classifier (Section 4.1). We show empirically this simple scheme can be sufficient to attain MEDoE's performance benefits. However, future work approaches to obtaining DoE classifiers could be investigated and tested. DoE classifiers that can be updated throughout the fine-tuning stage could be investigated. Furthermore, our experiments are in fully-observable environments, so future work could test MEDoE in partially observable environments and develop new approaches to learning DoE classifiers under partial observability.

REFERENCES

[1] Stefano V Albrecht, Filippos Christianos, and Lukas Schäfer. 2024. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches* (1 ed.). MIT Press, Cambridge, MA, USA.

[2] Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. 2022. VMAS: A Vectorized Multi-Agent Simulator for Collective Robot Learning. *The 16th International Symposium on Distributed Autonomous Robotic Systems* (2022).

[3] Michael Bowling. 2004. *Convergence and No-Regret in Multiagent Learning.* Technical Report. University of Alberta Libraries. https://doi.org/10.7939/R3ZS2KF41

[4] Michael Bowling and Manuela Veloso. 2001. Rational and Convergent Learning in Stochastic Games. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01, Vol. 2).* Morgan Kaufmann Publishers Inc., Seattle, WA, USA, 1021–1026.

[5] Yu-han Chang, Tracey Ho, and Leslie Pack Kaelbling. 2003. All Learning Is Local: Multi-agent Learning in Global Reward Games. In *Advances in Neural Information Processing Systems*, Vol. 16. MIT Press, Vancouver, Canada, 807–814.

[6] Caroline Claus and Craig Boutilier. 1998. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence (AAAI '98/IAAI '98).* American Association for Artificial Intelligence, USA, 746–752.

[7] Felipe Leno da Silva and Anna Helena Reali Costa. 2019. A Survey on Transfer Learning for Multiagent Reinforcement Learning Systems. *Journal of Artificial Intelligence Research* 64 (March 2019), 645–703. https://doi.org/10.1613/jair.1.11396

[8] Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. 2020. Emergent Complexity and Zero-Shot Transfer via Unsupervised Environment Design. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'21, Vol. 34).* Curran Associates Inc., Online Conference, 13049–13061.

[9] Elliot Fosong, Arrasy Rahman, Ignacio Carlucho, and Stefano V. Albrecht. 2022. Few-Shot Teamwork. In *Workshop on Ad Hoc Teamwork at IJCAI 2022.* arXiv, Vienna, Austria. https://doi.org/10.48550/arXiv.2207.09300 arXiv:2207.09300 [cs]

[10] John C. Harsanyi and Reinhard Selten. 1992. *A General Theory of Equilibrium Selection in Games* (2. print ed.). MIT Pr, Cambridge, Mass.

[11] Jeewon Jeon, Woojun Kim, Whiyoung Jung, and Youngchul Sung. 2022. MASER: Multi-agent Reinforcement Learning with Subgoals Generated from Experience Replay Buffer. *arXiv:2206.10607 [cs]* (June 2022). arXiv:2206.10607 [cs]

[12] Wenhao Li, Dan Qiao, Baoxiang Wang, Xiangfeng Wang, Bo Jin, and Hongyuan Zha. 2023. Semantically Aligned Task Decomposition in Multi-Agent Reinforcement Learning. arXiv:2305.10865 [cs]

[13] Siqi Liu, Guy Lever, Zhe Wang, Josh Merel, S. M. Ali Eslami, Daniel Hennes, Wojciech M. Czarnecki, Yuval Tassa, Shayegan Omidshafiei, Abbas Abdolmaleki, Noah Y. Siegel, Leonard Hasenclever, Luke Marris, Saran Tunyasuvunakool, H. Francis Song, Markus Wulfmeier, Paul Muller, Tuomas Haarnoja, Brendan Tracey, Karl Tuyls, Thore Graepel, and Nicolas Heess. 2022. From Motor Control to Team Play in Simulated Humanoid Football. *Science Robotics* 7, 69 (Aug. 2022), eabo0235. https://doi.org/10.1126/scirobotics.abo0235

[14] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. 2022. A Survey of Ad Hoc Teamwork Research. In *Multi-Agent Systems: 19th European Conference.* Springer-Verlag, Düsseldorf, Germany, 275–293. https://doi.org/10.1007/978-3-031-20614-6_16

[15] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. *Journal of Machine Learning Research* 21, 181 (2020), 1–50.

[16] Hadi Nekoei, Akilesh Badrinaaraayanan, Aaron Courville, and Sarath Chandar. 2021. Continuous Coordination as a Realistic Scenario for Lifelong Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139).* PMLR, Online Conference. arXiv:2103.03216

[17] Frans A. Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs* (1 ed.). Springer International Publishing, Cham, Switzerland. https://doi.org/10.1007/978-3-319-28929-8

[18] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep Exploration via Bootstrapped DQN. *arXiv:1602.04621 [cs, stat]* (July 2016). arXiv:1602.04621 [cs, stat]

[19] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V. Albrecht. 2019. Dealing with Non-Stationarity in Multi-Agent Deep Reinforcement Learning. *arXiv:1906.04737 [cs, stat]* (June 2019). arXiv:1906.04737 [cs, stat]

[20] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. 2021. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks.* Online Conference. arXiv:2006.07869 [cs, stat]

[21] Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. 2022. Evolving Curricula with Regret-Based Environment Design. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. PMLR, Baltimore, MD, USA, 17473–17498.

[22] Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. 2021. Federated Reinforcement Learning: Techniques, Applications, and Open Challenges. *Intelligence & Robotics* 1, 1 (Oct. 2021), 18–57. https://doi.org/10.20517/ir.2021.02

[23] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning.* PMLR, 4295–4304.

[24] David Rother, Thomas Weisswange, and Jan Peters. 2023. Disentangling Interaction Using MaximumEntropy Reinforcement Learning in Multi-Agent Systems. In *European Conference on Artificial Intelligence.*

[25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv:1707.06347 [cs]* (Aug. 2017). https://doi.org/10.48550/arXiv.1707.06347 arXiv:1707.06347 [cs]

[26] Peter Stone, Gal A Kaminka, Sarit Kraus, and Jeffrey S Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence.* AAAI Press, Atlanta, GA, USA, 1504–1509. https://doi.org/10.5555/2898607.2898847

[27] Kefan Su, Siyuan Zhou, Chuang Gan, Xiangjun Wang, and Zongqing Lu. 2022. MA2QL: A Minimalist Approach to Fully Decentralized Multi-Agent Reinforcement Learning. *arXiv:2209.08244 [cs]* (Sept. 2022). arXiv:2209.08244 [cs]

[28] Xuting Tang, Jia Xu, and Shusen Wang. 2022. Transferable Multi-Agent Reinforcement Learning with Dynamic Participating Agents. *arXiv:2208.02424 [cs]* (Aug. 2022). arXiv:2208.02424 [cs]

[29] Adam Taylor, Ivana Dusparic, Maxime Gueriau, and Siobhan Clarke. 2019. Parallel Transfer Learning in Multi-Agent Systems: What, When and How to Transfer?. In *2019 International Joint Conference on Neural Networks.* IEEE, Budapest, Hungary, 1–8. https://doi.org/10.1109/IJCNN.2019.8851784

[30] Dhruva Tirumala, Alexandre Galashov, Hyeonwoo Noh, Leonard Hasenclever, Razvan Pascanu, Jonathan Schwarz, Guillaume Desjardins, Wojciech Marian Czarnecki, Arun Ahuja, Yee Whye Teh, and Nicolas Heess. 2020. Behavior Priors for Efficient Reinforcement Learning. *arXiv:2010.14274 [cs]* (Oct. 2020). arXiv:2010.14274 [cs]

[31] Peter Vrancx, Yann-Michaël De Hauwere, and Ann Nowé. 2011. Transfer Learning for Multi-Agent Coordination. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, Vol. 1. SciTePress - Science and Technology Publications, Rome, Italy, 263–272. https://doi.org/10.5220/0003185602630272

[32] Rundong Wang, Longtao Zheng, Wei Qiu, Bowei He, Bo An, Zinovi Rabinovich, Yujing Hu, Yingfeng Chen, Tangjie Lv, and Changjie Fan. 2023. Towards Skilled Population Curriculum for Multi-Agent Reinforcement Learning. arXiv:2302.03429 [cs]

[33] Rose E. Wang, Sarah A. Wu, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. 2020. Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration. *arXiv:2003.11778 [cs]* (July 2020). arXiv:2003.11778 [cs]

[34] Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. 2020. ROMA: Multi-agent Reinforcement Learning with Emergent Roles. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, Online Conference, 9876–9886.

[35] Weixun Wang, Tianpei Yang, Yong Liu, Jianye Hao, Xiaotian Hao, Yujing Hu, Yingfeng Chen, Changjie Fan, and Yang Gao. 2020. From Few to More: Large-scale Dynamic Multiagent Curriculum Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. Online Conference, 7293–7300. https://doi.org/10.1609/aaai.v34i05.6221 arXiv:1909.02790

[36] Ermo Wei and Sean Luke. 2016. Lenient Learning in Independent-Learner Stochastic Cooperative Games. *Journal of Machine Learning Research* 17, 84 (2016), 1–42.

[37] D.H. Wolpert and W.G. Macready. 1997. No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation* 1, 1 (April 1997), 67–82. https://doi.org/10.1109/4235.585893

[38] Tianpei Yang, Weixun Wang, Hongyao Tang, Jianye Hao, Zhaopeng Meng, Hangyu Mao, Dong Li, Wulong Liu, Yingfeng Chen, Yujing Hu, Changjie Fan, and Chengwei Zhang. 2021. An Efficient Transfer Learning Framework for Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., Online Conference.

[39] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and YI WU. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., New Orleans, USA, 24611–24624. arXiv:2103.01955