

# Zero-Shot Coordination in Ad Hoc Teams with Generalized Policy Improvement and Difference Rewards

Rupal Nigam

University of Illinois

Urbana-Champaign

Champaign, IL, United States

rupaln2@illinois.edu

Mikihsia Yuasa

University of Illinois

Urbana-Champaign

Champaign, IL, United States

myuasa2@illinois.edu

Niket Parikh

University of Illinois

Urbana-Champaign

Champaign, IL, United States

niketnp2@illinois.edu

Hamid Osooli

University of Illinois

Urbana-Champaign

Champaign, IL, United States

hosooli2@illinois.edu

Jacob Heglund

University of Illinois

Urbana-Champaign

Champaign, IL, United States

jheglun2@illinois.edu

Huy T. Tran

University of Illinois

Urbana-Champaign

Champaign, IL, United States

huytran1@illinois.edu

## ABSTRACT

Real-world multi-agent systems may require ad hoc teaming, where an agent must coordinate with previously unseen teammates in a zero-shot manner. Prior work either selects a pretrained policy based on an inferred model of the new teammates or pretrains a single policy that is robust to potential teammates. Instead, we propose to *dynamically leverage all pretrained policies* through two key ideas—generalized policy improvement and difference rewards—for efficient and effective knowledge transfer between different teams. We empirically demonstrate that our algorithm, Generalized Policy improvement for Ad hoc Teaming (GPAT), successfully enables zero-shot transfer to new teams in three simulated environments and demonstrate our algorithm in a real-world multi-robot setting.

## KEYWORDS

ad hoc teaming; zero-shot coordination; reinforcement learning

### ACM Reference Format:

Rupal Nigam, Niket Parikh, Hamid Osooli, Mikihsia Yuasa, Jacob Heglund, and Huy T. Tran. 2026. Zero-Shot Coordination in Ad Hoc Teams with Generalized Policy Improvement and Difference Rewards. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 9 pages. <https://doi.org/10.65109/TNEX7143>

## 1 INTRODUCTION

Ad hoc teaming (AHT), where an autonomous agent must coordinate with other unknown agents [27], is an open challenge for multi-agent systems. Consider a search-and-rescue mission where robots are deployed from different organizations and expected to coordinate with each other on the fly—these robots may have different biases in how they achieve an objective (e.g., risky vs. risk-averse) or have different capabilities (e.g., sensing vs. manipulation). Adapting to such differences would enable agents to effectively and autonomously complete tasks where the team is unknown prior to

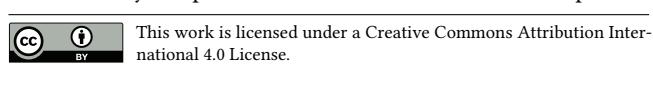
deployment. We therefore focus on zero-shot coordination (ZSC) for AHT, where the controlled agent, or the learner, is able to pretrain with various teams but then must coordinate with a new team with no online learning [14].

Type-based approaches are common for AHT, where the learner pretrains with a set of teammate types, then infers the best pretrained policy to use with the new team at test time [5, 12, 37]. However, these methods struggle to handle new teams not seen during pretraining and require online inference of the new team type. An alternative approach pretrains a learner that is robust to new teams through careful generation of diverse pretraining teams [11, 35, 38]. However, these methods often require large training populations, may suffer from overfitting, and can struggle to generalize to out-of-distribution teammates [32]. Generating many training teammates may also be infeasible for real-world applications, for example, due to computational constraints.

We address these challenges through two key ideas. Our first idea is to leverage a library of pretrained learner policies, but instead of choosing one at test time based on the inferred team type, we dynamically leverage the whole library with *no online inference or learning*. We specifically use a *generalized policy improvement* (GPI) policy to select from pretrained policies given the current state, motivated by the fact that GPI can guarantee improvement over library policies [2]. However, this guarantee is only valid when the dynamics are constant and the reward function differs—for AHT, dynamics now change due to new teammate behaviors, while the reward stays the same. Our second idea is then to use *difference rewards* to define the value functions of pretrained policies used by a GPI policy. Difference rewards address the multi-agent credit assignment problem by approximating the contribution of an individual agent to a team reward [20]. We use this idea to reduce the impact of the distribution shift induced by new teammates.

We integrate these ideas into an algorithm for ZSC in AHT and empirically demonstrate the benefits of our method. We summarize our contributions as follows:

- (1) we propose an algorithm for ZSC in AHT based on GPI with difference rewards (Section 5),
- (2) we empirically demonstrate the benefits of our method relative to baselines in three simulated environments and demonstrate its use in a multi-robot system (Section 6).

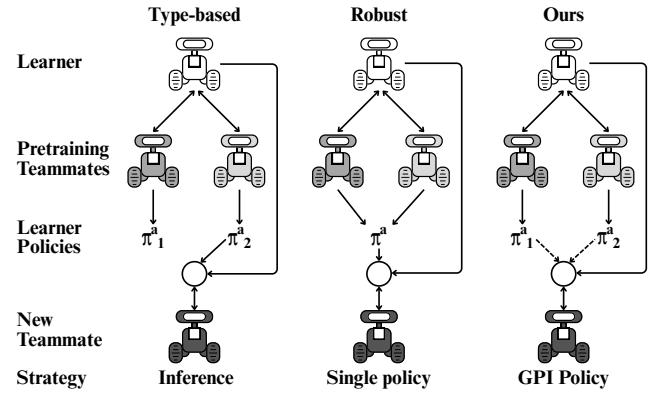


## 2 RELATED WORK

Type-based approaches to AHT use a pretrained library of policies, where the learner selects an appropriate response based on inferred teammate behavior. For instance, [5] updates a prior over the pretrained policies through online inference with the new teammate to determine the most likely teammate type, and acts using the corresponding learner policy. Similarly, [12] and [18] infer a teammate’s policy using a similarity metric and select a complementary learner policy to coordinate in Team Space Fortress. [36] leverages Gibbs sampling to update a distribution over possible learner policies in Hanabi and similarly use this distribution to select a pretrained policy. [37] extends these methods by employing a mixture-of-experts approach to mix pretrained policies in the Overcooked environment, rather than selecting a single policy. Their method first identifies behaviors through unsupervised clustering of previously collected data, trains learner policies for each behavior, and then uses online samples to update belief weights over each policy for a mixture-of-experts model. However, these methods require observations with the unseen teammate to appropriately infer the best response from the pretrained library, and thus are not truly zero-shot. Many of them also implement a single pretrained policy, limiting their ability to combine pretrained skills at execution.

Another class of methods focuses on training a single robust policy by generating a diverse training pool of teammates for ZSC. [11] achieves this by breaking symmetries inherent in the task through random relabeling of the teammate’s states and actions. [28] generates a diverse set of training teammates by using different random seeds and checkpoints during the training of agents. [35] expands this concept by searching over the reward space to construct a training pool. [38] enhances diversity in training teammates using an entropy bonus and employing a prioritization strategy to select training teammates. Lastly, [13] ensures diversity in teammate policies by considering intrinsic rewards with random navigation. However, these methods can also be prone to overfitting and may struggle to generalize to teammates outside of the training pool [32]. Additionally, many of these methods require large training pools, resulting in high computational costs. Finally, it may not be possible to define pretraining teammates in certain real-world settings, due to, for example, limited resources and access to robot platforms. Instead, it may be more practical in certain settings for a few representative teammates to be given to the learner.

We illustrate these two common approaches for AHT in Figure 1, alongside our proposed approach. Other AHT approaches are complementary to this paper. [9] and [23] address AHT with the additional challenge of partial observability, where teammate actions and environment rewards are unobservable, so the learner updates its belief prior over the library using online observations. [15] investigates the benefit of communication between teammates, while in this work we assume no communication between agents. [16] investigates the adaptation ability of AHT approaches in a few-shot, rather than zero-shot, coordination setting. [22] leverages graph structures to handle ad hoc teams where agents can enter and leave the team. [30] considers extending AHT to settings where multiple learners are present. We assume fixed teams with a single learner. [19] considers AHT with humans, but focus on the benefits of incorporating explainable AI. Finally, self-play (SP) methods [29]



**Figure 1: Common methods for AHT either select a pretrained policy based on the inferred behavior of the teammate (Type-based) or train a single policy that is robust to potential teammates by generating a diverse training pool (Robust). We instead dynamically select which policy to use at every timestep without inference for ZSC.**

have been successful in zero-sum settings [26], which are a type of AHT problem, but are ill-suited for cooperative AHT because they cannot coordinate effectively with non-SP agents [7, 11].

## 3 PRELIMINARIES

### 3.1 Multi-agent Reinforcement Learning

We model our problem as a multi-agent Markov decision process (MMDP) defined by a tuple  $\langle \mathcal{S}, \mathcal{N}, \{\mathcal{A}^i\}_{i \in \mathcal{N}}, p, r, \gamma \rangle$  [6]. Here,  $\mathcal{S}$  is the state space,  $\mathcal{N}$  is the set of agents,  $\mathcal{A}^i$  is the action space of agent  $i$ , and  $\gamma \in [0, 1)$  is the discount factor. Let  $\mathcal{A} := \times_{i \in \mathcal{N}} \mathcal{A}^i$  be the joint action space. Then  $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  is the state transition function and  $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$  is the team reward function. At time step  $t$ , each agent  $i \in \mathcal{N}$  executes an action  $a_t^i \in \mathcal{A}^i$  given the current state  $s_t \in \mathcal{S}$ , after which the system transitions to state  $s_{t+1} \in \mathcal{S}$  and the team receives reward  $r(s_t, a_t, s_{t+1})$ , where  $a_t \in \mathcal{A}$  is the joint action. Let  $\pi^i : \mathcal{S} \times \mathcal{A}^i \rightarrow [0, 1]$  be an individual policy for agent  $i$  and  $\pi(a|s) = \prod_{i \in \mathcal{N}} \pi^i(a^i|s)$  be the resulting joint policy of all agents. The performance of a joint policy  $\pi$  can be described by its action-value function,

$$Q^\pi(s, a) := \mathbb{E}_{p, \pi} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 = s, a_0 = a \right]. \quad (1)$$

### 3.2 Generalized Policy Improvement

Consider a set of source tasks  $\mathcal{R} = \{r_i\}_{i=1}^n$ , where each task  $r_i \in \mathcal{R}$  is defined as a linear reward,

$$r_i(s, a, s') = \phi(s, a, s')^\top \mathbf{w}_i, \quad (2)$$

where  $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}^d$  is a function mapping to  $d$  features and  $\mathbf{w}_i \in \mathbb{R}^d$  is a weight vector specifying preferences over features. Following [2], define the successor features (SFs) of a policy  $\pi$  as,

$$\psi^\pi(s, a) := \mathbb{E}_{p, \pi} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t, s_{t+1}) \mid s_0 = s, a_0 = a \right]. \quad (3)$$

Then, the action-value function of  $\pi$  on task  $r_i$ ,  $Q_{r_i}^\pi(s, a)$ , is,

$$Q_{r_i}^\pi(s, a) = \psi^\pi(s, a)^\top \mathbf{w}_i. \quad (4)$$

Now assume that we pretrain an agent on the set of source tasks  $\mathcal{R}$  to generate a set of optimal policies  $\Pi = \{\pi_i^*\}_{i=1}^n$ , where  $\pi_i^*$  is an optimal policy for task  $r_i \in \mathcal{R}$ . Given a new target task  $r_{n+1} = \phi(s, a, s')^\top \mathbf{w}_{n+1}$ , we can use a GPI policy,  $\pi'$ , defined as,

$$\pi'(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \max_{\pi \in \Pi} Q_{r_{n+1}}^\pi(s, a), \quad (5)$$

to perform no worse than any policy in  $\Pi$  on this task. If we compute the set of SFs associated with each policy in  $\Pi$ ,  $\Psi = \{\psi^{\pi_i^*}\}_{i=1}^n$ , we can compute the set  $\{Q_{r_{n+1}}^{\pi_i^*}\}_{i=1}^n$  using Equation (4) and implement  $\pi'$  with no additional learning on the new target task. If additional learning is allowed, we can use  $\pi'$  to quickly optimize a policy for  $r_{n+1}$ , for example using SFQL (Algorithm 3 in [3]). We use GPI as a framework for ZSC in AHT.

### 3.3 Difference Rewards

In cooperative MARL, multiple agents interact within a shared environment to achieve a common goal, represented as a shared, team reward. A core challenge in this setting is the multi-agent credit assignment problem: determining which agent(s) were responsible for the reward resulting from their collective actions. Difference rewards offer a solution to this by approximating an individual agent's contribution to the team reward [8, 20]. Instead of a team reward signal, each agent computes their individual difference reward and optimizes a policy to maximize its expected return. More formally, the difference reward for agent  $i$ ,  $\Delta r^i$ , is defined as,

$$\Delta r^i(s, a, s') = r(s, a, s') - \mathbb{E}_{b^i \sim \pi^i} \left[ r \left( s, \langle a^{-i}, b^i \rangle, s' \right) \right], \quad (6)$$

where  $a^{-i}$  is the joint action of all agents other than  $i$ . We extend the concept of difference rewards to the AHT setting so the learner can overcome the credit assignment problem.

## 4 PROBLEM FORMULATION

We modify the general formulation of MMDPs for AHT as follows. Let  $a \in \mathcal{N}$  be the *learner* (i.e., the agent whose policy we aim to optimize) and  $\mathcal{N}_u = \mathcal{N} \setminus \{a\}$  be the complementary set of all teammates (i.e., uncontrolled agents). We assume that each teammate follows a fixed policy, which is unknown to the learner. Teammate policies may be suboptimal with respect to the team reward  $r$  and the ad hoc team considered due to, e.g., the teammates being trained for a different task or with different teammates, or being humans and having inherent biases towards different goals. We formally define this model as an ad hoc MMDP.

**Definition 1** (Ad Hoc MMDP). An ad hoc MMDP is defined by a tuple  $M = \langle \mathcal{S}, \mathcal{N}, a, \{\mathcal{A}^i\}_{i \in \mathcal{N}}, p, r, \{\pi^i\}_{i \in \mathcal{N}_u}, \gamma \rangle$ , where  $a \in \mathcal{N}$  is the learner,  $\mathcal{N}_u = \mathcal{N} \setminus \{a\}$  is the complementary set of teammates, and  $\pi^i : \mathcal{S} \times \mathcal{A}^i \rightarrow [0, 1]$  is the fixed policy of teammate  $i$ .

We assume that the reward function,  $r$ , is non-negative. Note that any bounded reward can be transformed into a non-negative reward because scalar addition renders the reward to be policy invariant. We refer to an ad hoc MMDP  $M$  as an ad hoc team.

The performance of a learner policy  $\pi^a$  in ad hoc team  $M$  can be described by its action-value function,

$$Q^{\pi^a, \pi^{-a}}(s, a^a) := \mathbb{E}_{p, \pi^a, \pi^{-a}} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 = s, a_0^a = a^a \right], \quad (7)$$

where  $\pi^{-a}$  is the joint policy of all teammates induced by  $\{\pi^i\}_{i \in \mathcal{N}_u}$ . Our objective is to compute an optimal policy,  $\pi^{a^*}$ , which satisfies,

$$Q^{\pi^{a^*}, \pi^{-a}}(s, a^a) := \max_{\pi^a} Q^{\pi^a, \pi^{-a}}(s, a^a), \quad (8)$$

for all  $s \in \mathcal{S}$  and  $a^a \in \mathcal{A}^a$ . Optimizing a learner policy for an ad hoc team  $M$  is equivalent to solving a single-agent Markov decision process with a transition function  $\tilde{p}$  that captures the impact of teammate policies  $\pi^{-a}$ .

Assume we are given a partially specified ad hoc team  $M_{\setminus \mathcal{N}_u} = \langle \mathcal{S}, \cdot, a, \mathcal{A}^a, p, r, \cdot, \gamma \rangle$  and  $m$  possible ad hoc teammates  $\{\langle \mathcal{A}^i, \pi^i \rangle\}_{i=1}^m$ . Then let  $\mathcal{M}$  be the set of possible ad hoc teams induced by those teammates. We formalize our ZSC in AHT problem as follows.

**Problem 1** (Zero-shot Coordination for AHT). Let  $\mathcal{M}_0 = \{M_i\}_{i=1}^n \subseteq \mathcal{M}$  be a given set of source ad hoc teams with which the learner can pretrain. Our objective is to synthesize an optimal learner policy  $\pi_{n+1}^{a^*}$  for a new ad hoc team  $M_{n+1} \in \mathcal{M} \setminus \mathcal{M}_0$  by leveraging pretraining with  $\mathcal{M}_0$  but with no online learning with the new team  $M_{n+1}$ .

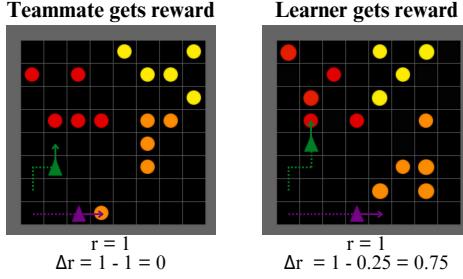
## 5 OUR METHOD

### 5.1 Key Ideas: Generalized Policy Improvement and Difference Rewards

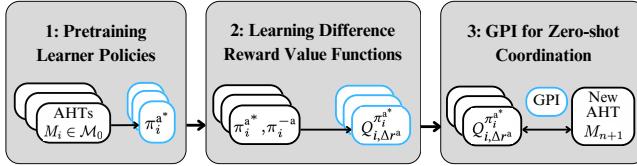
We address the AHT problem defined in Problem 1 through two key ideas. First, we use a GPI policy of the form in Equation (5) to dynamically leverage a library of pretrained learner policies to coordinate with a new ad hoc team with no online learning. By dynamic, we mean that GPI allows the learner to use different pretrained policies throughout the execution of a single episode, rather than being restricted to only using the best-matching pretrained policy (as in type-based methods) or a single robust policy (as in robust pretraining AHT methods). This ability to dynamically leverage policies is important, for example, in scenarios where a learner must use multiple pretrained skills to complete a task. Furthermore, this approach does not require online inference.

We are also motivated by the fact that a GPI policy guarantees improvement over each library policy in single-agent zero-shot transfer settings where dynamics are fixed and rewards are changed [3]. However, in our setting, new ad hoc teammates induce new dynamics due to their (potentially) different policies, while our team rewards are fixed. These new dynamics make the set of pretrained SFs no longer valid, which prevents instant evaluation of the value functions for the new task that GPI requires. To ensure policy improvement, one would need to perform policy evaluation for each pretrained policy with respect to the new ad hoc team, which requires many online samples.

We address this issue through our second idea—instead of having GPI operate over value functions of pretrained policies evaluated with respect to the team reward, we have it operate over value functions with respect to the learner's difference rewards.



**Figure 2: Difference rewards allow the learner to estimate its own contribution to the team reward.** We illustrate this in a foraging example where collecting any object gives a team reward of 1. (left) the teammate action results in the reward and (right) the learner action results in the reward. While the team reward is the same in both scenarios,  $\Delta r$  reflects the learner’s contribution. We use this idea to reduce the effect of the distribution shift induced by new teammates.



**Figure 3: Our method, GPAT, addresses ZSC in AHT through 3 steps:** (1) pretraining learner policies, (2) learning difference reward value functions to address the distribution shift induced by new teammates, and (3) dynamically leveraging this library at every time step by applying a GPI policy.

Note that these value functions still assume teammate dynamics associated with the ad hoc teams used during pretraining; that is, we do not correct them to account for the new ad hoc team and therefore policy improvement is still not guaranteed. However, we hypothesize that evaluating with respect to the learner’s difference rewards emphasizes contributions of the learner towards the team reward (while correspondingly de-emphasizing contributions of teammates), which will then reduce the impact of the distribution shift induced by the new ad hoc team on the actions selected by the GPI policy. This idea is illustrated in Figure 2.

## 5.2 An Algorithm for ZSC in AHT

Based on the ideas presented in Section 5.1, we propose an algorithm, GPI for Ad Hoc Teaming (GPAT), to address ZSC in AHT. Our algorithm is composed of three primary steps, visualized in Figure 3 and discussed below. Pseudocode is provided in Algorithm 1. We consider two settings, one where the team reward  $r$  can be modeled as a linear reward and the more general case of any team reward  $r$ . For the linear reward setting, we assume the features  $\phi$  are fixed and heuristically defined such that  $r$  can be modeled using Equation (2) through a weight vector  $w$ . Our algorithm can be extended to incorporate feature and weight learning, as in [4, 10].

---

### Algorithm 1 Generalized Policy Improvement for Ad Hoc Teaming

---

**Step 1:** Pretraining learner policies

**Require:**  $\Pi^{-a} \equiv \{\pi_1^{-a}, \dots, \pi_n^{-a}\}$

```

1: for team  $\in 1, \dots, i$  do
2:   if linear reward then
3:     Optimize  $\pi_i^a$  with  $\pi_i^{-a}$  using SFQL or SFDQN
4:   else if general reward then
5:     Optimize  $\pi_i^a$  with  $\pi_i^{-a}$  using any RL algorithm

```

**Step 2:** Learning difference reward value functions

**Require:**  $\Pi^a \equiv \{\pi_1^a, \dots, \pi_n^a\}, \Pi^{-a}, r(s, a), \phi(s, a)$

```

1: for team  $\in 1, \dots, i$  do
2:   for timestep  $\in 1, \dots, T_{DR}$  do
3:      $s \leftarrow \text{env.reset}()$ 
4:     while not done do
5:        $a \leftarrow \pi_i^a, \pi_i^{-a}$ 
6:        $s', r \leftarrow \text{env.step}(a)$ 
7:        $\Delta r \leftarrow \text{compute_dr}(r, a^{-a})$  ▷ Equation (6)
8:       if linear reward then
9:          $\mathcal{D} \leftarrow \Delta r, \phi(s, a)$ 
10:      else if general reward then
11:         $\delta \leftarrow \Delta r + \gamma Q_{i,\Delta r}(s', \pi^a(s')) - Q_{i,\Delta r}(s, a^a)$ 
12:         $\theta \leftarrow \theta + \alpha \delta \nabla_\theta Q_{i,\Delta r}(s, a^a)$ 
13:      if linear reward then
14:         $w_{i,\Delta r} \leftarrow \text{least_squares}(\mathcal{D})$  ▷ Equation (2)

```

**Step 3:** GPI for zero-shot coordination

**Require:**  $Q_{\Delta r^a} \equiv \{Q_{1,\Delta r^a}, \dots, Q_{n,\Delta r^a}\}, \pi_{n+1}^{-a}$

```

1:  $s \leftarrow \text{env.reset}()$ 
2: while not done do
3:    $a^a \leftarrow \text{argmax}_b \max_i Q_{i,\Delta r^a}(s, b)$  ▷ Equation (11)
4:    $a \leftarrow (a^a, \pi^{-a}(s))$ 
5:    $s, r \leftarrow \text{env.step}(a)$ 

```

---

**Step 1: Pretraining Learner Policies.** Given a set of source ad hoc teams  $\mathcal{M}_0$ , we first optimize a learner policy for each source ad hoc team. The output of this step is a library of learner policies  $\Pi^a = \{\pi_i^{a^*}\}_{i=1}^n$ , where  $\pi_i^{a^*}$  is the optimal learner policy for ad hoc team  $M_i \in \mathcal{M}_0$ . Any single-agent RL algorithm can be used for this step. In this work, we used Q-Learning with SFs (SFQL, Algorithm 3 in [3]) for linear reward settings. This process produces a set of optimal learner SFs  $\Psi^a = \{\psi_i^{\pi_i^{a^*}, \pi_i^{-a}}\}_{i=1}^n$ , where the learner SFs for learner policy  $\pi^a$  in ad hoc team  $M_i$  are defined as,

$$\psi^{\pi^a, \pi_i^{-a}}(s, a^a) = \mathbb{E}_{p, \pi^a, \pi_i^{-a}} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t, s_{t+1}) \mid s_0 = s, a_0^a = a^a \right], \quad (9)$$

where  $\pi_i^{-a}$  is the joint policy of all teammates in ad hoc team  $M_i$ . Following Equation (9), we use these learner SFs to define corresponding action-value functions for policies in  $\Pi^a$  as,

$$Q^{\pi^a, \pi_i^{-a}}(s, a^a) = \psi^{\pi^a, \pi_i^{-a}}(s, a^a)^\top w. \quad (10)$$

We refer to  $\psi^{\pi^a, \pi_i^{-a}}$  as  $\psi_i^{\pi^a}$  and  $Q^{\pi^a, \pi_i^{-a}}$  as  $Q_i^{\pi^a}$  hereafter to simplify notation, where  $\psi_i^{\pi^a}$  and  $Q_i^{\pi^a}$  are the SFs and action-value function for learner policy  $\pi^a$  in ad hoc team  $M_i \in \mathcal{M}$ . We used PPO [25] for general reward settings.

**Step 2: Learning Difference Reward Value Functions.** Given the library of pretrained learner policies  $\Pi^a$ , we now perform policy

evaluation with respect to the learner’s difference reward  $\Delta r^a$ , rather than the team reward  $r$ . Because the learner policies are deterministic, we assume a uniform learner policy when computing the difference rewards as in [33]. The output of this step is a set  $Q_{\Delta r^a}^a = \{Q_{i,\Delta r^a}^a(s, a^a)\}_{i=1}^n$ , where  $Q_{i,\Delta r^a}^a$  is the value function of policy  $\pi^a$  with respect to the learner’s difference reward in  $M_i \in \mathcal{M}$ .

*Linear Reward Setting.* We model the learner’s difference reward as  $\Delta r^a(s, a, s') = \phi(s, a, s')^\top w_{\Delta r^a}$ . Given this model,  $\Psi^a$ , and rollouts from the optimal learners in each ad hoc team  $M_i \in \mathcal{M}_0$ , we can now approximate  $Q_{\Delta r^a}^a$  by simply estimating  $w_{\Delta r^a}$  with linear regression and using Equation (10). We used Step 2 from Algorithm 1 for this work, and show that a sufficiently accurate  $w_{\Delta r^a}$  can be learned in **as few as 10 episodes** in Section 6. Note that this policy evaluation step can be performed during Step 1 using the same sampled experiences when using linear rewards—we simply separate this step for presentation purposes.

*General Reward Setting.* Any policy evaluation method can be used to estimate  $Q_{\Delta r^a}^a$  in the general reward setting. We use a TD-learning approach outlined in Step 2 from Algorithm 1 for this work, which is a simplified version of fitted Q-iteration (FQI) algorithms [17, 24]. This process can be computationally expensive, but allows one to model a broader class of rewards.

*Step 3: GPI for Zero-shot Coordination.* We finally use a GPI policy for ZSC in a target (new) ad hoc team  $M_{n+1}$ . Given  $Q_{\Delta r^a}^a$ , we define the GPI policy for the learner as,

$$\pi^a(s) \in \operatorname{argmax}_{a^a \in \mathcal{A}^a} \max_{i \in \{1, \dots, n\}} Q_{i,\Delta r^a}^{\pi^a}(s, a^a). \quad (11)$$

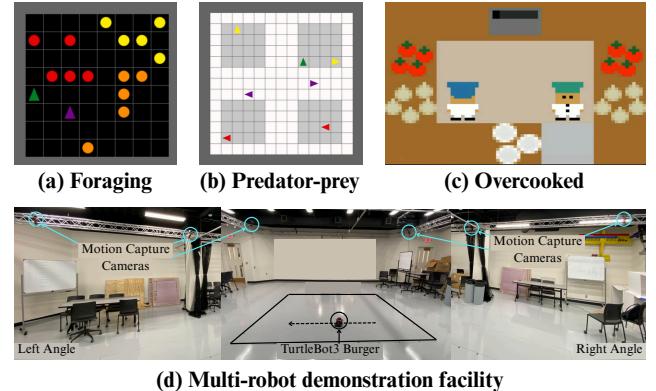
## 6 EXPERIMENTS

### 6.1 Experimental Setup

*Environments.* We empirically demonstrate GPAT’s performance in a multi-agent foraging environment inspired by [4, 9], a multi-agent predator-prey environment as in [9, 34], and Overcooked [7, 31]. The environments are illustrated in Figure 4. We assume linear rewards for all environments, though we also consider a general reward setting for foraging within our ablation study.

The foraging environment has 2 agents in a team (one learner and one teammate) that aim to collect 3 types of objects (red, orange, and yellow) in an  $8 \times 8$  grid. We define environment features  $\phi$  as  $1 \times 3$  vectors, where each element of  $\phi$  is the number of objects of a given type that are collected in a state transition. The team reward is then defined using  $w = [1, 1, 1]$ . Following [4], our state representations are agent-centric and toroidal, such that the agent is always in the upper left corner and the grid is wrapped around the edges of the environment. The representation has 5 channels: one for each object type, one for the teammate, and one for walls. We cluster each object type in separate quadrants of the grid. Within each cluster, objects spawn in random locations at the start of each episode, and agents spawn in the lower left quadrant.

The predator-prey environment has 3 agents in a team (predators) that aim to capture 4 prey in a  $13 \times 13$  grid. Prey randomly move within their shaded regions. We consider easy prey (yellow), which can be captured by a single predator, and hard prey (red), which must be captured by two predators. We define environment



**Figure 4: Multi-agent environments used in our experiments, where the learner agent is green.** (a) Circles represent different types of objects agents aim to collect. (b) Agents aim to capture easy (yellow) and hard (red) prey. (c) Agents must coordinate to cook and serve a tomato onion soup. (d) Our multi-robot demonstration uses TurtleBot3 Burgers.

features and state representations similar to those in the foraging environment, with the team reward defined using  $w = [1, 1, 1, 1]$ .

The Overcooked environment has 2 agents in a team that aim to cook and deliver soups in a  $2 \times 3$  room—we use the Cramped Room layout, modified to include an additional soup ingredient, as in [35], to increase coordination complexity (frequent interactions due to the small room size) and task complexity, as recommended in [32]. The agents must place an onion in the pot, place a tomato in the pot, start up the stove, wait for the soup to finish cooking, plate the soup, and finally deliver the soup. Following prior work [35], we define environment features to consider 5 rewarding events (potting onion, potting tomato, picking up a dish, picking up the soup, delivering the soup) and define the team reward as  $w = [3, 3, 3, 5, 20]$ . We use the default fully-observable state representation from [7].

*Teammate Policies.* For the foraging environment, we consider two source ad hoc teams in each experiment, where each experiment considers different combinations of source and new target ad hoc teams. We design teammate policies to have different preferences for which object types (colors) they collect—the ideal learner collects the objects ignored by its teammate. We optimize teammate policies using SFQL [4]. Table 1 defines our source and new target ad hoc teams, including statistics for the average number of object types each teammate collects in an episode. We also include statistics for the resulting pretrained learners and an oracle learner that is able to train with the new target ad hoc team.

For the predator-prey environment, we consider two source ad hoc teams, where the teammates use heuristic greedy policies designed so that the agents have preferences for which prey they target—similar to the foraging environment, the ideal learner targets prey ignored by its teammates.

The teammate in Overcooked follows heuristic greedy policies designed such that the agent has preferences for performing specific tasks for the recipe. Thus, the ideal learner performs the complementary tasks so the soup recipe can be completed and delivered.

**Table 1: Foraging experiment descriptions. The reward weights and objects collected vectors correspond to [red, orange, yellow]. Skills from pretrained learners that are useful for the new AHT, and used by the oracle, are bolded. We design experiments to test 3 scenarios with differing percentages of useful pretrained skills.**

		Source AHT 1		Source AHT 2		New AHT	
		Teammate	Learner	Teammate	Learner	Teammate	Oracle
<i>Experiment 1</i> <i>(50% useful prior skills)</i>	Reward weights	[1.0, -0.5, -0.5]	[1.0, 1.0, 1.0]	[-0.5, 1.0, -0.5]	[1.0, 1.0, 1.0]	[-0.5, -0.5, 1.0]	[1.0, 1.0, 1.0]
	Objects collected	[4.9, 0.0, 0.0]	[0.1, 5.0, 4.9]	[0.0, 4.6, 0.0]	[5.0, 0.2, 4.9]	[0.1, 0.1, 4.8]	[4.9, 4.9, 0.1]
	Preferred objects	red	<b>orange</b> , yellow	orange	<b>red</b> , yellow	yellow	<b>red, orange</b>
<i>Experiment 2</i> <i>(100% useful prior skills)</i>	Reward weights	[0.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[1.0, 0.0, 1.0]	[1.0, 1.0, 1.0]	[-0.5, -0.5, 1.0]	[1.0, 1.0, 1.0]
	Objects collected	[0.0, 4.9, 4.5]	[5.0, 0.1, 0.5]	[5.0, 0.0, 4.8]	[0.0, 5.0, 0.2]	[0.1, 0.1, 4.8]	[4.9, 4.9, 0.1]
	Preferred objects	orange, yellow	<b>red</b>	red, yellow	<b>orange</b>	yellow	<b>red, orange</b>
<i>Experiment 3</i> <i>(0% useful prior skills)</i>	Reward weights	[0.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[1.0, 0.0, 1.0]	[1.0, 1.0, 1.0]	[1.0, 1.0, 0.0]	[1.0, 1.0, 1.0]
	Objects collected	[0.0, 4.9, 4.5]	[5.0, 0.1, 0.5]	[5.0, 0.0, 4.8]	[0.0, 5.0, 0.2]	[4.8, 2.8, 0.2]	[0.1, 2.2, 4.7]
	Preferred objects	orange, yellow	red	red, yellow	orange	red, orange	<b>yellow</b>

*Learner Policies.* For the foraging and predator-prey environments, the learner policies were trained using SFQL [4] with simple multi-layer perceptron networks with two hidden layers of sizes 64 and 128. We used  $\epsilon$ -greedy exploration during training with  $\epsilon = 0.1$ . Updates were done with batch sizes of 10 with a discount factor of  $\gamma = 0.95$  and a learning rate of  $\alpha = 3 \times 10^{-5}$ . All policies were trained for 2,500,000 timesteps. For the more complex Overcooked environment, we implemented an SF-DQN algorithm using the Stable Baselines3 DQN algorithm [21]. We used the default hyperparameters, except for  $\gamma = 0.95$ , and 2,500,000 timesteps. Code for all experiments and environments is publicly available<sup>1</sup>.

*Baselines.* We implement GPAT with linear rewards in our experiments (unless otherwise noted) and learn  $w_{\Delta r}$  with 10 episodes, based on preliminary experiments ranging from 10-100 evaluation episodes. We compare GPAT to the following baselines:

- *Oracle*: We train a learner from scratch (using SFQL or SFDQN) with the new target AHT to represent an oracle learner.
- *Robust*: We pretrain a learner with all source AHTs to represent robust ZSC methods like [11, 35, 38], where we randomly sample the training team as in [35]. We train robust learners (using SFQL or SFDQN) for the total timesteps used to train the entire learner policy libraries used by other methods. We assume the source AHTs are given to the learner, so we do not consider the source AHT generation process.
- *PLASTIC*: We use the best pretrained learner policy from the library  $\Pi^a$  to represent a best-case scenario for type-based methods like [5, 36, 37]. We skip the online inference phase and simply use the best pretrained policy to mimic a setting where inference is already performed and correct.
- *MESH*: We implement a zero-shot version of [37], which combines pretrained policies from  $\Pi^a$  using a mixture of experts model. We use a uniform mixture since we do not allow online learning.

## 6.2 Results

Our results show inter-quartile mean (IQM) evaluation returns for the learner with the new target partner with 95% confidence

intervals (CIs), following the recommendations from [1]. Higher IQM scores are better. The CIs are estimated using percentile bootstrap with stratified sampling with 1,000 bootstrap resamples. We saw similar trends when analyzing mean and median results (not shown), and thus only show IQM results because it is more robust to outliers than the mean and has less bias than the median [1].

*Foraging.* We evaluate GPAT in the three experiments shown in Table 1. Experiment 1 requires the learner to leverage one of the two skills learned from each pretrained policy in order to optimally coordinate with the new teammate. Experiment 2 requires the learner to leverage all skills learned during pretraining. Experiment 3 contains no relevant skills in the learner’s pretrained policies. Table 2 shows results for these experiments. We see that GPAT (with a linear reward) outperforms all baselines (other than the oracle) for Experiments 1 and 2, likely because it is able to effectively use both pretrained skills. In comparison, the Robust baseline is likely struggling to generalize its pretrained skills to the new out-of-distribution teammate. The PLASTIC baseline is limited to using the best-matching pretrained policy and thus can only implement one of the needed skills. However, GPAT performs worse than both baselines in Experiment 3, likely due to the learner’s pretrained library not containing the skill needed to coordinate with the new teammate (i.e., collecting yellow objects). In principle, GPAT could perform similarly to PLASTIC, since it’s library contains the best-matching policy used by PLASTIC—however, we see that GPAT is outperformed by PLASTIC, indicating that it occasionally picks the worse policy, likely due to inaccuracies incurred by the distribution shift induced by the new teammate. Note that our ablation results (shown in Table 3) show that GPAT with general rewards performs similarly to our baselines in Experiment 3. We discuss this point more in the ablation discussion. Overall, these experiments suggest that GPAT can effectively achieve ZSC when its library has at least some relevant skills, but can struggle when there are no relevant skills in the library.

*Predator-Prey.* In the predator-prey experiment, the first pretrained learner policy learns to target the top-right and bottom-right prey, while the second pretrained learner policy learns to target the top-left and bottom-left prey. To coordinate optimally with the new AHT, the learner must target the top-left and bottom-right prey,

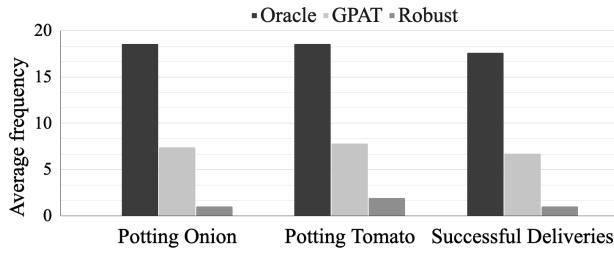
<sup>1</sup><https://github.com/apurl1/gpat>

**Table 2: Experiment results, showing IQM returns for GPAT compared to baselines [5, 11, 35], measured using 1000 episodes with 10 replicates and 95% confidence intervals. The percentage optimality with respect to the oracle is also reported in parentheses.**

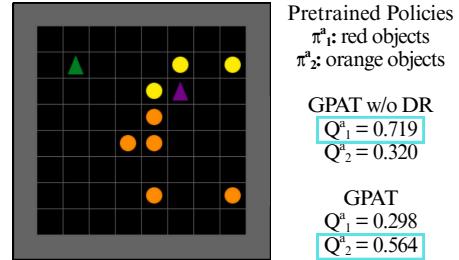
	Foraging			Predator-Prey	Overcooked
	Experiment 1	Experiment 2	Experiment 3		
Oracle	8.087 ± 0.011	8.082 ± 0.011	8.091 ± 0.012	2.547 ± 0.008	607.76 ± 0.735
<b>GPAT (ours)</b>	<b>7.755 ± 0.014(95.9%)</b>	<b>7.635 ± 0.019(94.5%)</b>	<b>5.641 ± 0.024(69.7%)</b>	<b>2.202 ± 0.011(86.5%)</b>	<b>218.07 ± 5.401(35.9%)</b>
Robust	5.998 ± 0.018(74.2%)	6.590 ± 0.016(81.5%)	<b>6.573 ± 0.016(81.2%)</b>	2.077 ± 0.011(81.5%)	33.05 ± 2.078(5.4%)
PLASTIC	6.438 ± 0.010(79.6%)	6.235 ± 0.012(77.1%)	6.238 ± 0.013(77.1%)	2.096 ± 0.008(82.3%)	9.69 ± 0.328(1.6%)
MESH	4.654 ± 0.032(57.5%)	5.485 ± 0.030(67.9%)	5.930 ± 0.024(73.3%)	1.569 ± 0.012(61.6%)	0.0 ± 0.0(0.0%)

**Table 3: Ablation results in the foraging environment, showing IQMs returns for GPAT with a linear reward (ours), general reward (GR), and without difference rewards (w/o DR), measured using 1000 episodes with 10 replicates and 95% confidence intervals. We also report the percentage optimality and the percentage each pretrained policy is used.**

	Experiment 1			Experiment 2			Experiment 3		
	Return	% $\pi_1$	% $\pi_2$	Return	% $\pi_1$	% $\pi_2$	Return	% $\pi_1$	% $\pi_2$
GPAT (ours)	<b>7.755 ± 0.014(95.9%)</b>	56.5%	43.5%	<b>7.635 ± 0.019(94.5%)</b>	50.1%	49.9%	5.641 ± 0.024(69.7%)	10.0%	90.0%
GPAT with GR	7.266 ± 0.022(89.8%)	56.2%	43.8%	7.492 ± 0.021(92.7%)	45.2%	54.8%	<b>6.282 ± 0.024(77.6%)</b>	5.6%	94.4%
GPAT w/o DR	7.066 ± 0.019(87.4%)	34.7%	65.3%	6.625 ± 0.019(82.0%)	72.4%	27.6%	6.206 ± 0.015(76.7%)	22.6%	77.4%



**Figure 5: The average number of times the learner performs various subtasks over an episode, measured using 1000 episodes with 10 replicates. The Robust learner struggles to pot onions and tomatoes, despite learning how to optimally perform both during pretraining.**



**Figure 6: Q-values for a state from Experiment 2 in the foraging environment, normalized by the maximum for each method. Without difference rewards, the learner is unable to switch policies once all red objects have been collected.**

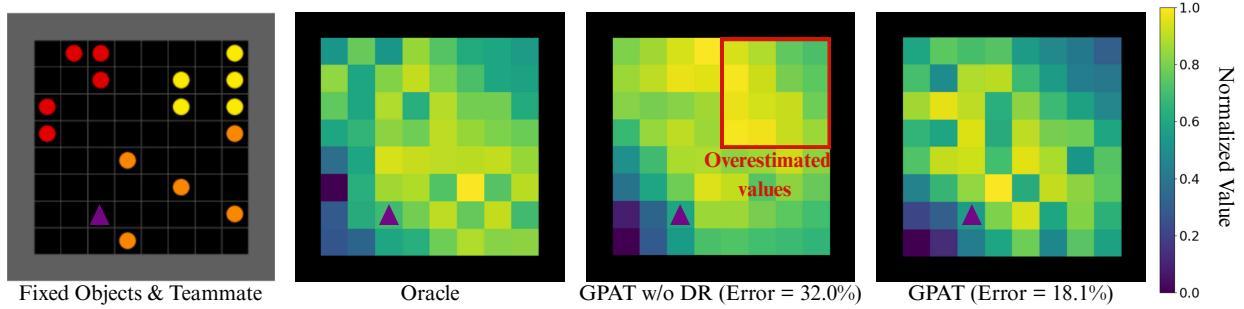
thus combining both pretrained skills. As shown in Table 2, we see that GPAT outperforms our baselines due to its ability to correctly use both pretrained policies as needed. Thus, GPAT can coordinate well with multiple teammates and handle dynamic environments.

**Overcooked.** In Overcooked, the first pretrained learner trains with an onion-preferring teammate to learn to pot tomatoes, cook the soup, plate the soup, and deliver the soup. The second pretrained learner trains with a tomato-preferring teammate to learn to pot onions, cook the soup, plate the soup, and deliver the soup. To coordinate optimally with the new dish-preferring teammate, the learner must pot both onions and tomatoes before cooking the soup. Table 2 shows that GPAT significantly outperforms our baselines. Compared to other environments, there is also a greater optimality gap between all methods and the oracle. We hypothesize that these trends are due to the sequential nature of the task, where it is

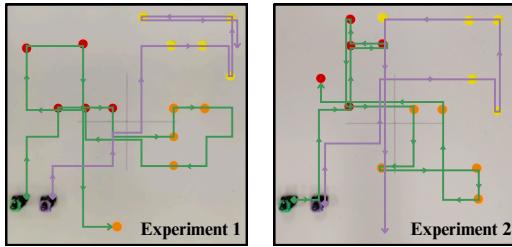
necessary to complete each step in the correct order, which thus requires more intelligent coordination than other environments.

We further investigate why the Robust learner struggles to perform well, despite converging to optimal returns during pretraining with library teammates. Figure 5 visualizes the number of times the learner pots onions or pots tomatoes, along with how many deliveries the team successfully made. We observe that the Robust learner does not perform the required subtasks often enough, indicating that it may have overfit to the training teammates and does not know how to optimally use those skills with the new teammate.

**Ablation Results.** Table 3 shows results from an ablation study for GPAT. We used PPO from Stable Baselines3 [21] to train learner policies with general rewards, and performed policy evaluation using TD-updates (see Algorithm 1) with 2500 episodes, based on preliminary experiments ranging from 100-5000 episodes. We see



**Figure 7:** Normalized value maps for a state (left) in the foraging environment for Experiment 2, where we show the value of the learner being in each grid position and the average percent error of the values relative to the oracle’s in parenthesis. Without difference rewards, the learner overestimates the value of being near yellow objects (the new teammate’s preferred object) and has a higher error than GPAT.



**Figure 8: Multi-robot demonstration of the foraging environment.** We overlay the trajectories taken by the robots for two experiments from Table 1 (green=learner, purple=teammate).

that GPAT with linear rewards outperforms general rewards in Experiments 1 and 2, likely due to easier reward learning. This, along with the much greater sample efficiency of learning  $w_{\Delta r}$  compared to learning  $Q_{\Delta r}$ , is why we use GPAT with linear rewards in the main results. However, GPAT with general rewards outperforms linear rewards in Experiment 3, where it performs similarly to the Robust and PLASTIC baselines. Recall that in Experiment 3, the pretraining library does not contain any relevant skills for the new teammate. Thus, we hypothesize that the noisier policy learned by general rewards is actually advantageous in this scenario. Table 3 also shows that removing difference rewards results in one pretrained policy being used much more than the other—ideally, we would expect the learner to roughly select each pretrained policy an equal number of times in Experiments 1 and 2. Figure 6 shows an example state where GPAT w/o DR is unable to switch between policies. Here, all red objects have been collected. Since the new teammate prefers yellow objects, the learner should switch to collecting orange (using  $Q_2^a$ ). However, GPAT without DR is unable to appropriately switch policies because it overestimates the value of  $\pi_1^a$ , since its value functions are with respect to the team reward and thus assume the teammate will collect the remaining objects. Furthermore, Figure 7 shows that removing difference rewards decreases performance, as it results in overestimated values near yellow objects (which is the new teammate’s preference) due to the value function being with respect to the team reward. We also see that, on average, GPAT w/o DR has a higher value error than GPAT.

Thus, difference rewards results in better aligned values despite the new dynamics induced by the new teammate.

### 6.3 Real World Multi-Robot Demonstration

We also demonstrate our method in a real-world multi-robot setting using Robotis Turtlebot3 Burgers in the foraging environment. We employ two TurtleBot3 Burger robots, each running a separate instance of the Robot Operating System (ROS) and maintaining its own identification parameters. A Qualysis motion capture system, with four passive markers attached to the robots, provides precise localization within a 12 ft.  $\times$  12 ft. grid-world. At every time step, we use GPAT to generate control commands, which are then transmitted to the robots through ROS. The trajectories the robots take are visualized in Figure 8. We observe the learner robot (green) collects the red and orange objects while the teammate robot (purple) collects the yellow objects, as expected. A video illustrating these real-world experiments is available online.<sup>2</sup>

## 7 CONCLUSION

In this work, we propose GPAT, a novel approach that leverages GPI and difference rewards for ZSC in AHT. In contrast to previous methods that use a single policy from a pretrained library with online inference or train a single policy robust to diverse teammates, we dynamically leverage the entire library of pretrained policies at every time step by applying a GPI policy. We address the impact of the distribution shift induced by new teammates on the GPI policy through difference rewards. We empirically demonstrate that this more exhaustive use of prior knowledge improves performance in three simulated environments relative to baselines. We also include a multi-robot demonstration of our method. Future directions include theoretical analysis, considering online adaptation, and considering partial observability.

## ACKNOWLEDGMENTS

This work was supported in part by ONR N00014-20-1-2249, a NASA grant awarded to the Illinois/NASA Space Grant Consortium, and a GAANN grant from the U.S. Department of Education.

<sup>2</sup>Robot video: <https://youtu.be/sm6Y-AakSgA>

## REFERENCES

- [1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G Bellemare. 2021. Deep Reinforcement Learning at the Edge of the Statistical Precipice. *Advances in Neural Information Processing Systems* (2021).
- [2] Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. 2018. Transfer in Deep Reinforcement Learning Using Successor Features and Generalised Policy Improvement. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 501–510. <https://proceedings.mlr.press/v80/barreto18a.html>
- [3] Andre Barreto, Will Dabney, Remi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. 2017. Successor Features for Transfer in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/350db081a661525235354dd3e19b8c05-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/350db081a661525235354dd3e19b8c05-Paper.pdf)
- [4] André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. 2020. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences* 117, 48 (Dec. 2020), 30079–30087. <https://doi.org/10.1073/pnas.1907370117> Publisher: Proceedings of the National Academy of Sciences.
- [5] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. 2017. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence* 242 (2017), 132–171. <https://doi.org/10.1016/j.artint.2016.10.005>
- [6] Craig Boutilier. 1996. Planning, Learning and Coordination in Multiagent Decision Processes. In *Proceedings of the Theoretical Aspects of Reasoning about Knowledge, TARK-96*.
- [7] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the Utility of Learning about Humans for Human-AI Coordination. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vol. 32. arXiv:1910.05789
- [8] Jacopo Castellini, Sam Devlin, Frans A. Oliehoek, and Rahul Savani. 2022. Difference Rewards Policy Gradients. *Neural Comput & Applic* (Nov. 2022). <https://doi.org/10.1007/s00521-022-07960-5>
- [9] Pengjie Gu, Mengchen Zhao, Jianye Hao, and Bo An. 2022. Online Ad Hoc Teamwork under Partial Observability. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=18Ys0-PzyPI>
- [10] Steven Hansen, Will Dabney, Andre Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. 2020. Fast Task Inference with Variational Intrinsic Successor Features. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=BJeAHkrYDS>
- [11] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. “Other-Play” for Zero-Shot Coordination. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 4399–4410. <https://proceedings.mlr.press/v119/hu20a.html> ISSN: 2640-3498.
- [12] Huo Li, Tianwei Ni, Siddharth Agrawal, Fan Jia, Suhas Raja, Yikang Gui, Dana Hughes, Michael Lewis, and Katia Sycara. 2021. Individualized Mutual Adaptation in Human-Agent Teams. *IEEE Transactions on Human-Machine Systems* 51, 6 (2021), 706–714. <https://doi.org/10.1109/THMS.2021.3107675>
- [13] Keane Lucas and Ross E Allen. 2022. Any-Play: An Intrinsic Augmentation for Zero-Shot Coordination. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 853–861.
- [14] Reuth Mirsky, Ignacio Carlculo, 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*, Dorothea Baumeister and Jörg Rothe (Eds.). Springer International Publishing, Cham, 275–293.
- [15] Reuth Mirsky, William Macke, Andy Wang, Harel Yedidsion, and Peter Stone. 2020. A Penny for Your Thoughts: The Value of Communication in Ad Hoc Teamwork. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 254–260. <https://doi.org/10.24963/ijcai.2020/36>
- [16] Hadi Nekoei, Xutong Zhao, Janarthanan Rajendran, Miao Liu, and Sarath Chandar. [n.d.]. Towards Few-shot Coordination: Revisiting Ad-hoc Teamplay Challenge In the Game of Hanabi. In *Proceedings of The 2nd Conference on Lifelong Learning Agents* (2023-11-20). PMLR, 861–877. <https://proceedings.mlr.press/v232/nekoei23b.html>
- [17] Gerhard Neumann and Jan Peters. 2008. Fitted Q-iteration by advantage weighted regression. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) (*NIPS’08*). Curran Associates Inc., Red Hook, NY, USA, 1177–1184.
- [18] Tianwei Ni, Huo Li, Siddharth Agrawal, Suhas Raja, Fan Jia, Yikang Gui, Dana Hughes, Michael Lewis, and Katia Sycara. 2021. Adaptive Agent Architecture for Real-time Human-Agent Teaming. In *Proceedings of AAAI ’21 W20: Workshop on Plan, Activity, and Intent Recognition (PAIR ’21)*.
- [19] Rohan Paleja, Muyleng Ghuy, Nadun R Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. 14.
- [20] Scott Proper and Kagan Turner. 2012. Modeling Difference Rewards for Multiagent Learning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*. 2.
- [21] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dorman. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. <http://jmlr.org/papers/v22/20-1364.html>
- [22] Arrasy Rahman, Niklas Hopner, Filippos Christanos, and Stefano V Albrecht. 2021. Towards Open Ad Hoc Teamwork Using Graph-based Policy Learning. In *Proceedings of the 38th International Conference on Machine Learning*. 11.
- [23] João G Ribeiro, Cassandro Martinho, Alberto Sardinha, and Francisco S Melo. 2022. Assisting unknown teammates in unknown tasks: Ad hoc teamwork under partial observability. *arXiv preprint arXiv:2201.03538* (2022).
- [24] Martin A. Riedmiller. 2005. Neural Fitted Q Iteration - First Experiences with a Data Efficient Neural Reinforcement Learning Method. In *European Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:6921329>
- [25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 <http://arxiv.org/abs/1707.06347>
- [26] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529, 7587 (2016), 484–489. <https://doi.org/10.1038/nature16961> arXiv:1610.00633
- [27] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. *Proceedings of the AAAI Conference on Artificial Intelligence* 24, 1 (Jul. 2010), 1504–1509. <https://doi.org/10.1609/aaai.v24i1.7529>
- [28] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. *Advances in Neural Information Processing Systems* 34 (2021), 14502–14515.
- [29] Gerald Tesauro. 1994. TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. 6, 2 (1994), 215–219. <https://doi.org/10.1162/neco.1994.6.2.215>
- [30] Caroline Wang, Muhammad Arrasy Rahman, Ishan Durugkar, Elad Liebman, and Peter Stone. 2025. N-agent ad hoc teamwork. *Advances in Neural Information Processing Systems* 37 (2025), 111832–111862.
- [31] 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. <https://doi.org/10.48550/arXiv.2003.11778> [cs].
- [32] Xihuai Wang\*, Shao Zhang\*, Wenhao Zhang, Wentao Dong, Jingxiao Chen, Ying Wen, and Weinan Zhang. 2024. Zsc-Eval: An Evaluation Toolkit and Benchmark for Multi-agent Zero-shot Coordination. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- [33] David H Wolpert and Kagan Turner. 2001. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 02n03 (2001), 265–279.
- [34] Dong Xing, Pengjie Gu, Qian Zheng, Xinrun Wang, Shanqi Liu, Longtao Zheng, Bo An, and Gang Pan. 2023. Controlling Type Confounding in Ad Hoc Teamwork with Instance-wise Teammate Feedback Rectification. In *Proceedings of the 40th International Conference on Machine Learning*.
- [35] Chao Yu, Jiaxuan Gao, Weilin Liu, Botian Xu, Hao Tang, Jiaqi Yang, Yu Wang, and Yi Wu. 2023. Learning Zero-Shot Cooperation with Humans, Assuming Humans Are Biased. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=TrwE8l9ajzs>
- [36] Jaleh Zand, Jack Parker-Holder, and Stephen J. Roberts. 2022. On-the-fly Strategy Adaptation for ad-hoc Agent Coordination. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (Virtual Event, New Zealand) (AAMAS ’22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1771–1773.
- [37] Michelle Zhao, Reid Simmons, and Henny Admoni. 2022. Coordination With Humans Via Strategy Matching. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 9116–9123. <https://doi.org/10.1109/IROS47612.2022.9982277> ISSN: 2153-0866.
- [38] Rui Zhao, Jimming Song, Yufeng Yuan, Hu Haifeng, Yang Gao, Yi Wu, Zhongqian Sun, and Yang Wei. 2022. Maximum Entropy Population-Based Training for Zero-Shot Human-AI Coordination. <https://doi.org/10.48550/arXiv.2112.11701> [cs].