Settling Decentralized Multi-Agent Coordinated Exploration by Novelty Sharing

Haobin Jiang[1], **Ziluo Ding**[1,2], **Zongqing Lu**[1†]
[1]School of Computer Science, Peking University
[2]Beijing Academy of Artificial Intelligence
{haobin.jiang, ziluo, zongqing.lu}@pku.edu.cn

Abstract

Exploration in decentralized cooperative multi-agent reinforcement learning faces two challenges. One is that the novelty of global states is unavailable, while the novelty of local observations is biased. The other is how agents can explore in a coordinated way. To address these challenges, we propose MACE, a simple yet effective multi-agent coordinated exploration method. By communicating only local novelty, agents can take into account other agents' local novelty to approximate the global novelty. Further, we newly introduce weighted mutual information to measure the influence of one agent's action on other agents' accumulated novelty. We convert it as an intrinsic reward *in hindsight* to encourage agents to exert more influence on other agents' exploration and boost coordinated exploration. Empirically, we show that MACE achieves superior performance in three multi-agent environments with sparse rewards.

Introduction

Recent progress in decentralized learning theories and algorithms for multi-agent reinforcement learning (MARL) (Zhang et al. 2018; de Witt et al. 2020; Jin et al. 2021; Daskalakis, Golowich, and Zhang 2022; Jiang and Lu 2022) makes it feasible to learn high-performant policies in a decentralized way for *cooperative* multi-agent tasks. However, one critical issue remains, *i.e.*, how to enable agents to effectively explore in a coordinated way under such a learning paradigm, especially for sparse-reward tasks where the environment rarely provides rewards.

One of the most popular exploration schemes in the single-agent setting is novelty-based exploration (Bellemare et al. 2016; Pathak et al. 2017; Burda et al. 2018b; Zhang et al. 2021b), where the agent is encouraged by well-designed intrinsic reward to visit *novel* states it rarely sees. However, things could be different when migrating to decentralized multi-agent settings, which leads to an unsolved problem: as only the local observation instead of the global state is available, how should each agent measure the novelty of the global state?

In decentralized settings, partial observability expands the discrepancy between each agent's local observation novelty and global state novelty, which makes the exploration merely based on local novelty highly unreliable (Wang et al. 2019a; Iqbal and Sha 2019). Fortunately, communication can help ease partial observability by providing extra information about other agents (Jiang and Lu 2018; Das et al. 2019; Wang et al. 2019b; Ding, Huang, and Lu 2020). However, unlimited communication may incur too much communication overhead, and it can indeed transform the decentralized setting into a centralized setting. Therefore, we resort to decentralized learning with *limited communication* to address such problems.

In addition to the challenge of novelty measurement, in cooperative tasks, agents must also acquire the ability to coordinate with each other to explore and achieve the final goal. Ideally, the optimal exploration strategy should consider others' observations and actions. Previous work (Wang et al. 2019a; Iqbal and Sha 2019; Liu et al. 2021) finds that independent exploration is not efficient and redundant exploration occurs. By coordination in exploration, we mean agents help other agents to achieve novel observations or reach novel states together through cooperation. In other words, an agent should be encouraged when its action enables other agents to reach more novel observations.

In this paper, we propose a *simple yet effective* **M**ulti-**A**gent **C**oordinated **E**xploration method, namely **MACE**. MACE introduces a *novelty-based* intrinsic reward and a *hindsight-based* intrinsic reward to enable coordinated exploration in decentralized cooperative tasks. Within the confines of limited communication, agents only share their local novelties (merely a floating point number) during training. Each agent leverages this shared information to approximate the global novelty, which serves as the novelty-based intrinsic reward. This approach aims to bridge the gap between the local novelty and the global novelty. Moreover, we encourage agents to exert more influence on others' explorations through the hindsight-based intrinsic reward, thereby boosting coordinated exploration. To this end, we measure the weighted mutual information (Guiasu 1977; Schaffernicht and Gross 2011) between the action of the agent and the accumulated novelty obtained thereafter by others given the local observation. The higher the weighted mutual information value, the higher the hindsight-based intrinsic reward the agent receives.

We evaluate MACE in three multi-agent environments:

†Corresponding author

GridWorld, Overcooked (Carroll et al. 2019), and SMAC (Samvelyan et al. 2019). All tasks in these environments are sparse-reward and hard to explore. The experimental results verify the effectiveness of MACE. Through ablation studies, we show that both the approximation to global novelty and the encouragement to influence other agents' exploration are indispensable in decentralized multi-agent exploration, and our newly employed weighted mutual information works significantly better than normal mutual information.

Preliminary

Decentralized learning. We consider an N-agent Markov decision process (MDP) $\mathcal{M} = \{S, \boldsymbol{O}, \boldsymbol{A}, P, R, \gamma\}$. Here, S represents the state space while $\boldsymbol{A}$ is the joint action space; the transition probability is defined by $P(s'|s, \boldsymbol{a})$. In decentralized learning, each agent has only access to its own local observation $o_i \in O_i$, rather than the global state, and learns an independent policy π_i to maximize the shared reward defined by R together with other agents. Notably, decentralized learning is more practical than centralized learning, owing to its better scalability, privacy, and security (Zhang, Yang, and Basar 2019).

Limited communication. We allow agents to communicate during the training phase. However, in order to enhance the practicality and adhere closely to the decentralized setting, we impose constraints on the bandwidth of the communication channel to reduce communication overhead (Foerster et al. 2016; Kim et al. 2018; Wang et al. 2020a). Specifically, the message sent by one agent at each step is confined to a floating point number. In this paper, we set agents to communicate their local novelties through this limited channel. Communication is not allowed during execution.

Note that this setting differs from centralized training and decentralized execution (CTDE) (Lowe et al. 2017; Foerster et al. 2018; Rashid et al. 2018), where agents can use unlimited extra information to ease training, such as other agents' observations and actions, or a centralized value function. Besides, our setting is not identical to *fully* decentralized learning (Tan 1997; de Witt et al. 2020; Jiang and Lu 2022), where communication is forbidden. On top of the fully decentralized learning algorithm, we will show that adding communication of novelty during training can enable coordinated exploration of agents to solve sparse-reward tasks.

Methodology

In this section, we present MACE addressing the challenges in decentralized multi-agent exploration. MACE follows the line of intrinsically motivated exploration (Yang et al. 2021) that designs intrinsic rewards r_{int} and trains agents via the shaped reward $r_s = r_{\text{ext}} + r_{\text{int}}$, where r_{ext} denotes the extrinsic reward given by the environment. MACE adopts the following two parts to design intrinsic rewards: 1) To obtain a more reliable novelty estimate as the novelty-based intrinsic reward, MACE uses the summation of all agents' novelty to approximate the global novelty. 2) To boost coordinated exploration, MACE further quantifies the influence of agents on the accumulated future novelty of other agents and converts it into the hindsight-based intrinsic reward. The weighted sum of these two parts is the final intrinsic reward used in MACE.

Approximation to Global Novelty

In decentralized training, if we only take into account the individual exploration of agent i, $u_t^i = \text{novelty}(o_{t+1}^i)$ can serve as the intrinsic reward to encourage agent i to take actions towards observations it seldom visits. When the observation space is discrete and small, such as the 2-dimension grid (x, y), we could directly record the number of times each observation that agent i has visited before and define $\text{novelty}(o) = 1/n(o)$ where $n(o)$ denotes the visit counts. If the observation space becomes large or continuous, methods designed for high-dimensional input such as pseudo-count (Bellemare et al. 2016), ICM (Pathak et al. 2017), and RND (Burda et al. 2018b) could be used to measure the novelty.

However, u_t^i only measures the local novelty of agent i. In the multi-agent environment, given the discrepancy between the local novelty and the global novelty, u_t^i may not be able to provide accurate and sufficient information for exploration. For example, we consider a two-agent environment where at timestep t, agent 1 is in an observation with low local novelty, and agent 2 is in an observation with high local novelty. From the global perspective, the two agents are in a novel state. However, from agent 1's perspective, it thinks that the observation is not novel and gives itself a low intrinsic reward, preventing it from further exploring the current novel state.

Due to the decentralized setting, the global novelty is not available to each agent. Therefore, we need a more appropriate intrinsic reward term than u_t^i to narrow the gap with the global novelty. Thanks to the limited communication, agents can exchange their local novelty u_t^i with each other at each timestep t. We propose a heuristic that uses the summation of all agents' local novelty as an approximation to the global novelty and as the novelty-based intrinsic reward:

$$r_{\text{nov}}^i(o_t^i, a_t^i) = \sum_j u_t^j. \tag{1}$$

With the introduction of other agents' novelty, we can avoid the aforementioned dilemma.

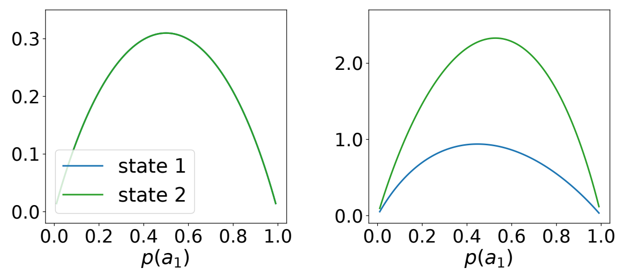
We admit that (1) still deviates from the global novelty in some cases. For example, agent 1 and agent 2 are in low-novelty observations while the global state is novel, which occurs when agent 1 and agent 2 seldom visit current observations *simultaneously*. Nevertheless, the gap with the global novelty cannot be closed entirely due to the limited information, and experimental results prove empirically that (1) works better than the local novelty u_t^i, as shown in the subsequent experimental section. One may argue that an alternative is to use the maximum of all agents' novelty as the intrinsic reward, but our preliminary experiment demonstrates that (1) works better.

Influence on Other Agents' Exploration

To boost coordinated exploration in multi-agent environments, each agent should consider its influence on other agents' exploration so that it could find some *critical states*

state 1 act	reward probability	**state 2** act	reward probability
a_1	$p(r=1 \mid a_1)=0.1$	a_1	$p(r=1 \mid a_1)=0.1$
	$p(r=5 \mid a_1)=0.8$		$p(r=5 \mid a_1)=0.8$
	$p(r=9 \mid a_1)=0.1$		$p(r=9 \mid a_1)=0.1$
a_2	$p(r=1 \mid a_2)=0.8$	a_2	$p(r=1 \mid a_2)=0.1$
	$p(r=5 \mid a_2)=0.1$		$p(r=5 \mid a_2)=0.1$
	$p(r=9 \mid a_2)=0.1$		$p(r=9 \mid a_2)=0.8$

Table 1: Action and reward probability of two illustrative states.



(a) MI (b) WMI

Figure 1: (a) Mutual information (MI) between action and reward in state 1 and state 2. (b) Weighted mutual information (WMI) between action and reward in state 1 and state 2.

(Yang et al. 2021). Critical states here mean that in these states, the action taken by one agent affects other agents' exploration progress, *e.g.*, one agent steps on a switch and thus opens a door that blocks another agent's way. So encouraging agents to explore these critical states would help them learn to cooperate effectively. We first discuss how to quantify one agent's influence on other agents' exploration.

Suppose there are two agents, agent i and agent j, in the environment. To estimate agent i's influence on agent j's exploration in a specific observation, we could use *mutual information*, a common measure used in MARL (Li et al. 2022), to quantify the dependence between agent i's action a_t^i and agent j's accumulated novelty $z_t^j = \sum_{t'=t} \gamma^{t'-t} u_{t'}^j$ given agent i's observation o_t^i:

$$I\left(A_t^i; Z_t^j | o_t^i\right) = \mathbb{E}_{a_t^i, z_t^j | o_t^i}\left[\log \frac{p(a_t^i, z_t^j | o_t^i)}{p(a_t^i | o_t^i) p(z_t^j | o_t^i)}\right].$$

Here we use agent j's accumulated novelty z_t^j instead of its immediate novelty u_t^j to measure the long-term dependence.

However, mutual information ignores the *magnitude* of agent j's accumulated novelty z_t^j. So it would give similar measurements for observation o_1^i where a_t^i is related to some low-value z_t^j, and observation o_2^i where a_t^i is related to some high-value z_t^j. To illustrate the statement intuitively, we devise two states with two actions and three different rewards, described in Table 1. Action and reward here can be seen as a_t^i and z_t^j respectively. As shown in Figure 1(a), with different $p(a_1)$, state 1 and state 2 always keep the same mutual information. But state 2 is more critical to coordinated exploration because the action (a_2) taken by agent i in state 2 can lead agent j to high accumulated novelty ($r=9$) more likely. Although agent i's action in state 1 has an influence on agent j's accumulated novelty to the same extent as that in state 2 measured by mutual information, it is more likely to result in lower accumulated novelty ($r=1$ or $r=5$). Therefore we need a more effective measure to estimate the influence of agent i's action a_t^i on agent j's accumulated novelty z_t^j, while taking into account the magnitude of z_t^j.

To this end, we newly introduce *weighted mutual information* (Guiasu 1977; Schaffernicht and Gross 2011) between agent i's action a_t^i and agent j's accumulated novelty z_t^j given agent i's observation o_t^i:

$$\omega I\left(A_t^i; Z_t^j | o_t^i\right) = \mathbb{E}_{a_t^i, z_t^j | o_t^i}\left[\omega(a_t^i, z_t^j) \log \frac{p(a_t^i, z_t^j | o_t^i)}{p(a_t^i | o_t^i) p(z_t^j | o_t^i)}\right] \tag{2}$$

where $\omega(\cdot,\cdot)$ denotes the weight placed on the pair of a_t^i and z_t^j. By introducing weights, pairs of a_t^i and z_t^j would have different informativeness. We set $\omega(a_t^i, z_t^j) = z_t^j$, meaning that relational mappings between a_t^i and higher z_t^j carry more significance than others. To illustrate how it works, Figure 1(b) shows weighted mutual information of state 1 and state 2 with different $p(a_1)$. We can see that the weighted mutual information of state 2 is always higher than that of state 1, consistent with what we expected. To summarize, (2) evaluates an observation o_t^i based not only on whether agent i's action has an influence on agent j's exploration, but also on whether agent i's action could lead to a high accumulated novelty of agent j.

Intrinsic Reward in Hindsight

To encourage each agent i to visit observations with high weighted mutual information, we define an intrinsic reward r_{wmi}^i of its observation o_t^i as:

$$r_{\text{wmi}}^i(o_t^i) = \sum_{j \neq i} r_{\text{wmi}}^{i \to j}(o_t^i) = \sum_{j \neq i} \omega I(A_t^i; Z_t^j | o_t^i). \tag{3}$$

$r_{\text{wmi}}^{i \to j}$ denotes the intrinsic reward given to agent i corresponding to its influence on agent j's exploration measured by weighted mutual information. Agent i's intrinsic reward r_{wmi}^i is the summation of all $r_{\text{wmi}}^{i \to j}$, representing its total influence on other agents' exploration. However, it is nontrivial to compute $r_{\text{wmi}}^{i \to j}$ according to (2), because it is an expectation over all actions and accumulated novelty. So we decompose the intrinsic reward (3) onto each action:

$$r_{\text{wmi}}^{i \to j}(o_t^i, a_t^i) = \mathbb{E}_{z_t^j | o_t^i, a_t^i}\left[z_t^j \log \frac{p(a_t^i, z_t^j | o_t^i)}{p(a_t^i | o_t^i) p(z_t^j | o_t^i)}\right]. \tag{4}$$

Further, we can continue to decompose (4) and get a hindsight-based intrinsic reward:

$$r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j) = z_t^j \log \frac{p(a_t^i, z_t^j | o_t^i)}{p(a_t^i|o_t^i)p(z_t^j|o_t^i)}. \quad (5)$$

Here $p(a_t^i|o_t^i)$ is the current policy $\pi^i(a_t^i|o_t^i)$ of agent i. With Bayesian rule $p(a_t^i|o_t^i, z_t^j) = \frac{p(a_t^i, z_t^j|o_t^i)}{p(z_t^j|o_t^i)}$, we can rewrite the logarithmic term in (5) and have:

$$r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j) = z_t^j \log \frac{p(a_t^i|o_t^i, z_t^j)}{\pi^i(a_t^i|o_t^i)}, \quad (6)$$

which is the form of the hindsight-based intrinsic reward we use in the paper. The term 'hindsight' reflects the difference between (6) and normal reward functions: (6) uses information obtained in future, *i.e.*, agent j's accumulated novelty z_t^j, which is not available until the end of the episode.

The logarithmic term in (6) is the *pointwise mutual information* between a_t^i and z_t^j. Pointwise mutual information measures the association between two random variables, commonly used in natural language processing (NLP). Therefore, (6) could be interpreted as encouraging action associative with the high accumulated novelty of agent j. If an a_t^i co-occurs with a high z_t^j at timestep t but there is no association between them, the logarithmic term in (6) will be around zero and agent i will not receive a high intrinsic reward, despite the high z_t^j.

Note that $r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j)$ keeps the following relationship with the original $r_{\text{wmi}}^{i\to j}(o_t^i)$:

$$\begin{aligned} r_{\text{wmi}}^{i\to j}(o_t^i) &= \mathbb{E}_{a_t^i|o_t^i}\left[r_{\text{wmi}}^{i\to j}(o_t^i, a_t^i)\right] \\ &= \mathbb{E}_{a_t^i, z_t^j|o_t^i}\left[r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j)\right], \end{aligned} \quad (7)$$

thus using $r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j)$ could be regarded as a Monte Carlo method for estimating $r_{\text{wmi}}^{i\to j}(o_t^i)$. Agent i needs agent j's accumulated novelty z_t^j and the posterior distribution $p(a_t^i|o_t^i, z_t^j)$ to calculate $r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j)$ at each timestep t. The former is computed at the end of the episode by accumulating agent j's novelty u_t^j that agent i obtained through communication. Note that this does not require additional communication, and each agent still communicates only local novelty at each timestep. The latter could be estimated from trajectory samples. To fulfill (7), samples used to estimate $p(a_t^i|o_t^i, z_t^j)$ and compute $r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j)$ should be on-policy because the expectation follows the distribution over z_t^j, determined by the current policies of all agents. So our proposed intrinsic reward is more suitable for *on-policy* reinforcement learning algorithms.

MACE

We combine the novelty-based intrinsic reward (1) and the hindsight-based intrinsic reward (6) to get the final shaped reward:

$$\begin{aligned} & r_{\text{s}}^i(o_t^i, a_t^i, \{z_t^j\}_{j\neq i}) \\ &\quad = r_{\text{ext}} + r_{\text{nov}}^i(o_t^i, a_t^i) + \lambda \sum_{j\neq i} r_{\text{hin}}^{i\to j}(o_t^i, a_t^i, z_t^j) \\ &\quad = r_{\text{ext}} + \sum_j u_t^j + \lambda \sum_{j\neq i} z_t^j \log \frac{p(a_t^i|o_t^i, z_t^j)}{\pi^i(a_t^i|o_t^i)}, \end{aligned} \quad (8)$$

where λ is a hyperparameter that denotes the weight of the hindsight-based intrinsic reward. In other words, λ controls the weight between encouraging agents to visit globally novel states and encouraging agents to influence other agents' exploration. Since the calculation of the hindsight-based intrinsic reward requires on-policy samples, we use independent PPO (IPPO) (Schulman et al. 2017; de Witt et al. 2020) as the base RL algorithm and train each agent i with shaped reward (8). To guarantee scalability, we additionally propose a scalable hindsight-based intrinsic reward using weighted mutual information between the agent's action and the summation of all other agents' accumulated novelty.

Related Work

Single-agent exploration. Advanced RL algorithms have been developed to improve exploration. Providing the agent with a manually designed intrinsic reward has been proven effective in environments with sparse rewards like Montezuma's Revenge (Brockman et al. 2016). Typically, the intrinsic reward is set to be the novelty of the state, *e.g.*, the inverse of the visit count: $r_{\text{int}}(s) = \text{novelty}(s) = 1/n(s)$, to encourage the agent to take action towards states it seldom visits. However, states in real problems are usually high-dimensional, meaning that $n(s)$ is impossible to count in most cases. Count-based methods solve this problem by introducing pseudo-count (Bellemare et al. 2016) or hashing to discretize states (Tang et al. 2017). Other methods measure novelty from different perspectives, including prediction error of transition model (Pathak et al. 2017; Burda et al. 2018a; Kim et al. 2019), prediction error of state features (Burda et al. 2018b), policy discrepancy (Flet-Berliac et al. 2020), state marginal matching (Lee et al. 2019), deviation of policy cover (Zhang et al. 2021a), uncertainty (Houthooft et al. 2016; Pathak, Gandhi, and Gupta 2019), and TD error of random reward (Ramesh et al. 2022). Recent work places an episodic restriction on intrinsic reward, where the intrinsic reward obtained by an agent visiting a repeated state within an episode will be reduced (Badia et al. 2019; Raileanu and Rocktäschel 2019; Zhang et al. 2021b; Henaff et al. 2022).

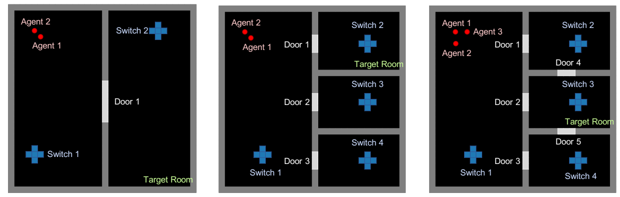
Multi-agent exploration. Exploration in multi-agent environments requires intrinsic reward that is different from that in single-agent environments. Iqbal and Sha (2019) proposed several types of intrinsic reward which take into consideration the novelty of agent i's observation from the perspective of agent j. EITI/EDTI (Wang et al. 2019a) focuses on encouraging the agent to states or observations where the agent influences other agents' transition or value function. EMC (Zheng et al. 2021) uses the summation of the prediction errors of local Q-functions as the shared intrinsic reward. MAVEN (Mahajan et al. 2019) improves multi-agent exploration by maximizing the mutual information between the trajectory and a latent variable, by which agents are encouraged to visit diverse trajectories. CMAE (Liu et al. 2021) combines the goal-based method with a state space dimension selection technique to adapt to the exponentially increased state space. MASER (Jeon et al. 2022) selects goals from the observation space instead of the state space.

However, these methods follow the CTDE setting, where unlimited extra information can be used to ease training. Some use QMIX (Rashid et al. 2018) as their backbone (Mahajan et al. 2019; Zheng et al. 2021; Liu et al. 2021; Jeon et al. 2022), while others require agents to share their local observations and actions (Iqbal and Sha 2019; Wang et al. 2019a). In contrast, MACE is built on top of decentralized learning algorithms and requires neither a centralized Q-function like QMIX nor the communication of observations and actions between agents. It only needs to pass a floating point number, *i.e.*, the local novelty, between agents, resulting in much less communication overhead than the methods mentioned above. Thus, comparison with these multi-agent exploration methods is out of the focus of this paper.

Decentralized multi-agent reinforcement learning. By virtue of the advantages of decentralized learning, *e.g.*, easy to implement, better scalability, and more robust (Jiang and Lu 2022), decentralized learning has attracted much attention from the MARL community. The convergence of decentralized learning was theoretically studied for cooperative games in networked settings (Zhang et al. 2018) and for fully decentralized (without communication) stochastic games in tabular cases (Jin et al. 2021; Daskalakis, Golowich, and Zhang 2022), laying the theoretical foundation for decentralized learning. de Witt et al. (2020); Papoudakis et al. (2021) showed the promising empirical performance of fully decentralized algorithms including IPPO and independent Q-learning (IQL) (Tan 1993) in several cooperative multi-agent benchmarks. Recently, Jiang and Lu (2022) proposed I2Q, a practical fully decentralized algorithm based on Q-learning for cooperative tasks, and proved the convergence of the optimal joint policy, yet limited to deterministic environments. However, the existing work does not take into consideration coordinated exploration and simply uses ϵ-greedy or sampling from the stochastic policy at individual agents. We take a step further to consider decentralized coordinated exploration and thus enable decentralized learning algorithms to solve sparse-reward tasks. As discussed before, our proposed hindsight-based intrinsic reward is more suitable for on-policy algorithms, thus we currently build MACE on IPPO. Combining MACE with off-policy decentralized algorithms like IQL or I2Q is left as future work.

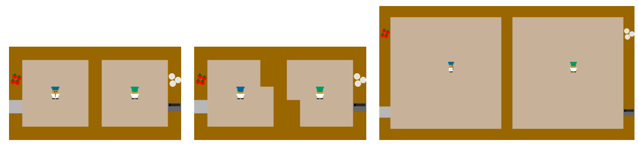
Experiments

In experiments, we evaluate MACE in three environments: GridWorld, Overcooked (Carroll et al. 2019), and SMAC (Samvelyan et al. 2019). We set all environments sparse-reward. Since we consider decentralized learning, agents in the experiments *do not share their parameters* and learn independently, following existing work (Jiang and Lu 2022).



(a) `Pass` (b) `SecretRoom` (c) `MultiRoom`

Figure 2: GridWorld: (a) `Pass`. (b) `SecretRoom`. (c) `MultiRoom`.



(a) `Base` (b) `Narrow` (c) `Large`

Figure 3: Overcooked: (a) `Base`. (b) `Narrow`. (c) `Large`.

GridWorld. We design three tasks in GridWorld including `Pass`, `SecretRoom`, and `MultiRoom`. `Pass` and `SecretRoom` reference tasks in Wang et al. (2019a) and Liu et al. (2021). In `MultiRoom`, the task extends to three agents. The goal of all tasks is that all agents enter the target room shown in Figure 2.

`Pass`: There are two agents in the 30×30 grid. *Door* 1 will open when any switch is occupied. To achieve the goal, one agent needs to reach *switch* 1 to open *door* 1 so that the other agent can enter the target room, then the latter agent needs to reach *switch* 2 to let the former agent come in. `SecretRoom`: There are two agents in the 30×30 grid. *Door* k will open when *switch* $k+1$ is occupied and all *doors* will open when *switch* 1 is occupied. Agents need to take the same steps as that in `Pass` to finish the task. `SecretRoom` is harder than `Pass` because there are three rooms on the right to explore but only one room is the target. `MultiRoom`: There are three agents in the 30×30 grid. Specifically, *door* 1 will open when *switch* 1 is occupied; *door* 3 will open when *switch* 2 is occupied; *door* 2 will open when *switch* 4 is occupied; *door* 4 and *door* 5 will open when *switch* 3 is occupied. More complicated coordinated exploration is required among the three agents.

The episode ends when all agents are in the target room, and each agent receives a +100 reward. Each agent observes its own location (x, y) and the open states of doors. These GridWorld tasks serve as didactic examples because the critical states in which exploration of agents interact with each other are obvious, namely the switch locations.

Overcooked. We design three tasks in Overcooked (Carroll et al. 2019): `Base`, `Narrow`, and `Large`. All tasks contain two agents, separated by an impassable kitchen counter as shown in Figure 3. Therefore, the two agents must cooperate to complete the task. When the soup is served, the episode ends, and each agent receives a +100 reward. Compared to `Base`, `Narrow` limits the area where items can be passed

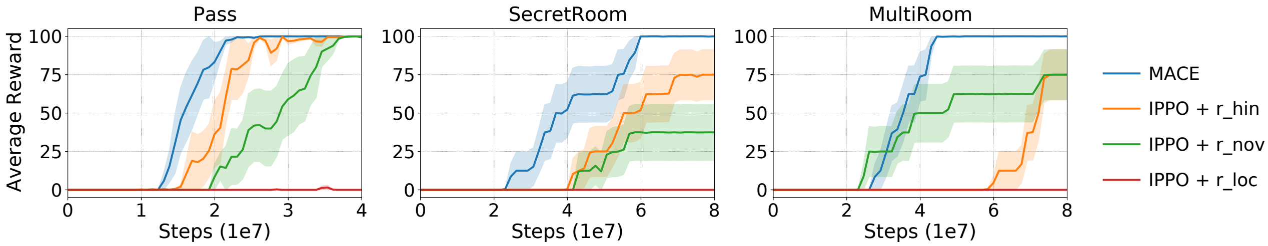


Figure 4: Learning curves of MACE compared with IPPO+r_loc, IPPO+r_nov, and IPPO+r_hin on three GridWorld tasks.

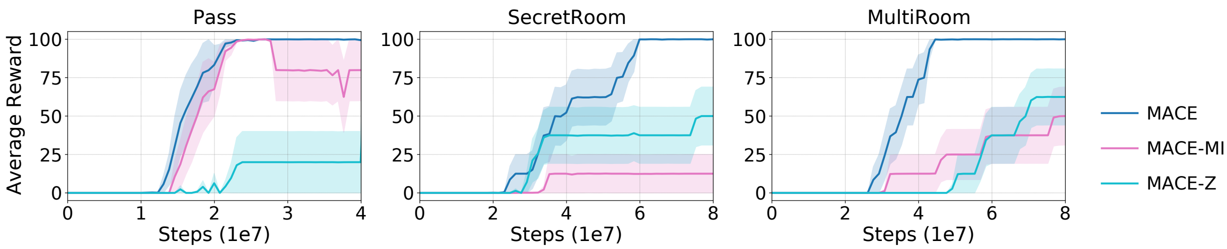


Figure 5: Learning curves of MACE compared with MACE-MI and MACE-Z on three GridWorld tasks.

to only the middle of the counter, and `Large` increases the size of the entire environment.

SMAC. We use three maps in SMAC (Samvelyan et al. 2019) 2.4.10: `2m_vs_1z`, `3m`, and `8m`, customized to be sparse-reward. Agents receive a +200 reward if they win the game. In `3m` and `8m`, agents also receive a +10 reward when one enemy dies so as to ease the task.

Implementation. For all tasks, we implement PPO leveraging GRU (Cho et al. 2014) as the policy and critic function. In GridWorld, given that the observation space is small and discrete, we use the inverse of visit counts as the novelty measurement and use a table to record each observation's visit count. Also, we use a table to record recent discretized accumulated novelty z_t^j and corresponding observation o_t^i and action a_t^i. Then we can estimate the posterior distribution $p(a_t^i|o_t^i, z_t^j)$ from the table. In Overcooked and SMAC, we use RND (Burda et al. 2018b) as the novelty measurement and use an MLP to fit the posterior distribution $p(a_t^i|o_t^i, z_t^j)$ via supervised learning.

GridWorld

We first verify the effectiveness of MACE in promoting coordinated exploration by ablation studies. We compare MACE with the following methods: a) IPPO+r_loc: agents are trained with $r_{\text{ext}} + u_t^i$, only taking into consideration the local novelty; b) IPPO+r_nov: agents are trained with $r_{\text{ext}} + r_{\text{nov}}^i$, exploring via approximated global novelty; c) IPPO+r_hin: agents are trained with $r_{\text{ext}} + u_t^i + \lambda r_{\text{hin}}^i$, exploring via local novelty and influence on other agents' exploration. λ here keeps the same as that used in MACE.

The results are shown in Figure 4. Each curve shows the mean reward of several runs with different random seeds (5 runs in `Pass`, 8 runs in `SecretRoom` and `MultiRoom`) and shaded regions indicate standard error. IPPO+r_loc is unable to solve any task because the local novelty is unreliable and insufficient for coordinated exploration. IPPO+r_nov performs better than IPPO+r_loc, indicating that taking into account the local novelty of other agents to approximate the global novelty is helpful for coordinated exploration. MACE achieves the best performance on all three tasks, suggesting that the hindsight-based intrinsic reward can further boost coordinated exploration by finding the critical states where the agent influences other agents' exploration. This can also be evidenced by the fact that IPPO+r_hin achieves higher rewards than IPPO+r_loc.

To further illustrate how the intrinsic rewards work, we visualize the novelty-based and hindsight-based intrinsic reward of agent 1 in the left room in `Pass`, averaging over 700 to 1000 PPO updates. The critical states consist of one agent stepping on one switch because it will open the middle door and allow the other agent to enter the target room and explore. As shown in Figure 8, agent 1 earns higher novelty-based intrinsic rewards at the bottom of the left room than at the top. The hindsight-based intrinsic reward can locate the critical states more accurately: agent 1 earns the highest hindsight-based intrinsic reward around *switch* 1.

The hindsight-based intrinsic reward (6) consists of two parts: z_t^j, the accumulated novelty of agent j, and a logarithmic term $\log \frac{p(a_t^i|o_t^i, z_t^j)}{\pi^i(a_t^i|o_t^i)}$. We test the effectiveness of the two parts separately to verify that none of them alone leads to MACE's high performance. MACE-MI replaces the hindsight-based intrinsic reward with the logarithmic term. We name it 'MI' because the expectation of this term equals the mutual information between a_t^i and z_t^j given o_t^i. The results in Figure 5 show that MACE-MI is less effective than MACE in all tasks, validating our claim that weighted mutual information is a more effective measure of the influence on other agent's exploration than mutual information. MACE-Z, replacing the hindsight-based intrinsic re-

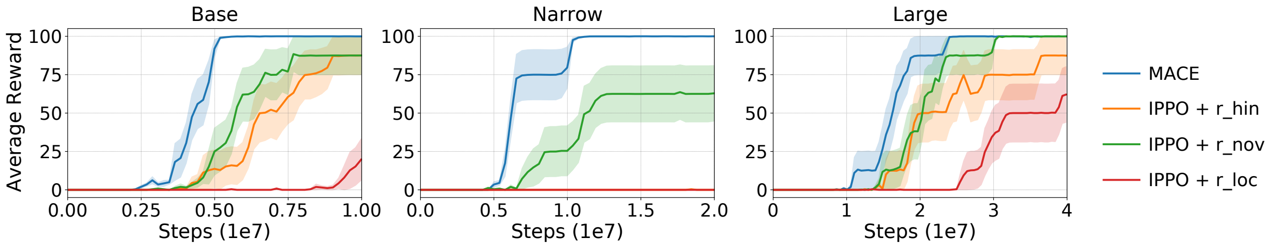


Figure 6: Learning curves of MACE compared with IPPO+r_loc, IPPO+r_nov, and IPPO+r_hin on three Overcooked tasks.

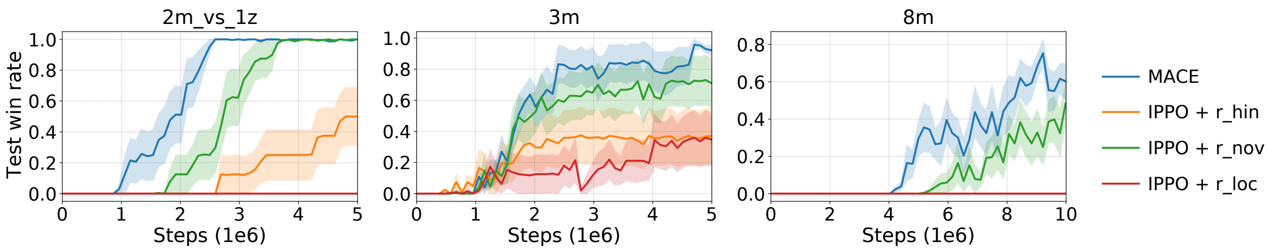
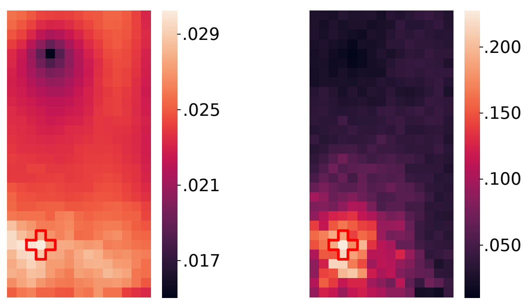


Figure 7: Learning curves of MACE compared with IPPO+r_loc, IPPO+r_nov, and IPPO+r_hin on three SMAC maps.



(a) novelty-based (b) hindsight-based

Figure 8: Visualization of the averaged (a) novelty-based and (b) hindsight-based intrinsic reward received by agent 1 at different positions in the left room. The red cross highlights the location of *switch* 1.

ward with z_t^j, also performs worse than MACE, indicating that utilizing other agents' accumulated novelty as the intrinsic reward, regardless of whether it is related to the agent's own actions, is ineffective.

Overcooked

We evauluate the performance of MACE in Overcooked (Carroll et al. 2019) and compare it with IPPO+r_loc, IPPO+r_nov, and IPPO+r_hin. The results are illustrated in Figure 6. Each curve shows the mean reward of 8 runs with different random seeds, and shaded regions indicate standard error. MACE outperforms others, proving that MACE also works in the high-dimensional state space where the novelty is calculated via RND (Burda et al. 2018b) and the posterior distribution $p(a_t^i|o_t^i, z_t^j)$ is learned via supervised learning. We also observe that IPPO+r_nov outperforms IPPO+r_hin in all tasks, especially in `Narrow`, suggesting that the novelty-based intrinsic reward may play a more critical role in such complicated tasks.

SMAC

We further examine MACE in more complex SMAC (Carroll et al. 2019) tasks and compare it with IPPO+r_loc, IPPO+r_nov, and IPPO+r_hin. The results are demonstrated in Figure 7 for three maps: `2m_vs_1z`, `3m`, and `8m`. Each curve shows the mean reward of 8 runs with different random seeds, and shaded regions indicate standard error. MACE learns faster or achieves a higher win rate than other baselines, by which we verify the effectiveness of MACE on sparse-reward tasks in such a high-dimensional complex environment.

Conclusion

We propose MACE to enable multi-agent coordinated exploration in decentralized learning with limited communication. MACE uses a novelty-based intrinsic reward and a hindsight-based intrinsic reward to guide exploration. The former is devised to narrow the gap between the local novelty and the unavailable global novelty. The latter is designed to find the critical states where one agent's action influences other agents' exploration, measured by the newly introduced weighted mutual information metric. Through empirical evaluation, we demonstrate the effectiveness of MACE in a variety of sparse-reward multi-agent tasks that need agents to explore cooperatively. In addition, we acknowledge that certain limitations exist, such as the need for a fully-connected communication network to share novelty among agents. Future work could explore ways to further reduce the number of necessary connections besides the bandwidth of communication channels.

Acknowledgments

This work was supported by NSF China under grant 62250068. The authors would like to thank the anonymous reviewers for their valuable comments.

References

Badia, A. P.; Sprechmann, P.; Vitvitskyi, A.; Guo, D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A.; et al. 2019. Never Give Up: Learning Directed Exploration Strategies. In *International Conference on Learning Representations*.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI Gym. *arXiv preprint arXiv:1606.01540*.

Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; and Efros, A. A. 2018a. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2018b. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

Cao, J.; Yuan, L.; Wang, J.; Zhang, S.; Zhang, C.; Yu, Y.; and Zhan, D.-C. 2021. Linda: Multi-agent local information decomposition for awareness of teammates. *arXiv preprint arXiv:2109.12508*.

Carroll, M.; Shah, R.; Ho, M. K.; Griffiths, T.; Seshia, S.; Abbeel, P.; and Dragan, A. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Das, A.; Gervet, T.; Romoff, J.; Batra, D.; Parikh, D.; Rabbat, M.; and Pineau, J. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*, 1538–1546. PMLR.

Daskalakis, C.; Golowich, N.; and Zhang, K. 2022. The Complexity of Markov Equilibrium in Stochastic Games. *arXiv preprint arXiv:2204.03991*.

de Witt, C. S.; Gupta, T.; Makoviichuk, D.; Makoviychuk, V.; Torr, P. H.; Sun, M.; and Whiteson, S. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*.

Ding, Z.; Huang, T.; and Lu, Z. 2020. Learning individually inferred communication for multi-agent cooperation. *Advances in Neural Information Processing Systems*, 33: 22069–22079.

Flet-Berliac, Y.; Ferret, J.; Pietquin, O.; Preux, P.; and Geist, M. 2020. Adversarially Guided Actor-Critic. In *International Conference on Learning Representations*.

Foerster, J.; Assael, I. A.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29.

Foerster, J. N.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual Multi-Agent Policy Gradients. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Guiasu, S. 1977. *Information Theory with Applications*. New York, NY: McGraw-Hill.

Henaff, M.; Raileanu, R.; Jiang, M.; and Rocktäschel, T. 2022. Exploration via Elliptical Episodic Bonuses. In *Advances in Neural Information Processing Systems*.

Houthooft, R.; Chen, X.; Duan, Y.; Schulman, J.; De Turck, F.; and Abbeel, P. 2016. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29.

Iqbal, S.; and Sha, F. 2019. Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. *arXiv preprint arXiv:1905.12127*.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *International Conference on Machine Learning (ICML)*.

Jeon, J.; Kim, W.; Jung, W.; and Sung, Y. 2022. Maser: Multi-agent reinforcement learning with subgoals generated from experience replay buffer. In *International Conference on Machine Learning*, 10041–10052. PMLR.

Jiang, J.; and Lu, Z. 2018. Learning Attentional Communication for Multi-Agent Cooperation. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiang, J.; and Lu, Z. 2021. The emergence of individuality. In *International Conference on Machine Learning*, 4992–5001. PMLR.

Jiang, J.; and Lu, Z. 2022. I2Q: A Fully Decentralized Q-Learning Algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jin, C.; Liu, Q.; Wang, Y.; and Yu, T. 2021. V-Learning - A Simple, Efficient, Decentralized Algorithm for Multiagent RL. *arXiv preprint arXiv:2110.14555*.

Kim, D.; Moon, S.; Hostallero, D.; Kang, W. J.; Lee, T.; Son, K.; and Yi, Y. 2018. Learning to Schedule Communication in Multi-agent Reinforcement Learning. In *International Conference on Learning Representations*.

Kim, H.; Kim, J.; Jeong, Y.; Levine, S.; and Song, H. O. 2019. EMI: Exploration with Mutual Information. In *International Conference on Machine Learning*, 3360–3369. PMLR.

Kim, W.; Jung, W.; Cho, M.; and Sung, Y. 2020. A maximum mutual information framework for multi-agent reinforcement learning. *arXiv preprint arXiv:2006.02732*.

Lee, L.; Eysenbach, B.; Parisotto, E.; Xing, E.; Levine, S.; and Salakhutdinov, R. 2019. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*.

Li, C.; Wang, T.; Wu, C.; Zhao, Q.; Yang, J.; and Zhang, C. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 3991–4002.

Li, P.; Tang, H.; Yang, T.; Hao, X.; Sang, T.; Zheng, Y.; Hao, J.; Taylor, M. E.; Tao, W.; and Wang, Z. 2022. PMIC: Improving Multi-Agent Reinforcement Learning with Progressive Mutual Information Collaboration. In *International Conference on Machine Learning*, 12979–12997. PMLR.

Liu, I.-J.; Jain, U.; Yeh, R. A.; and Schwing, A. 2021. Cooperative exploration for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 6826–6836. PMLR.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.

Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. Maven: Multi-agent variational exploration. *Advances in Neural Information Processing Systems*, 32.

Papoudakis, G.; Christianos, F.; Schäfer, L.; and Albrecht, S. V. 2021. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. arXiv:2006.07869.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

Pathak, D.; Gandhi, D.; and Gupta, A. 2019. Self-supervised exploration via disagreement. In *International conference on machine learning*, 5062–5071. PMLR.

Raileanu, R.; and Rocktäschel, T. 2019. RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments. In *International Conference on Learning Representations*.

Ramesh, A.; Kirsch, L.; van Steenkiste, S.; and Schmidhuber, J. 2022. Exploring through Random Curiosity with General Value Functions. *arXiv preprint arXiv:2211.10282*.

Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International conference on machine learning*, 4295–4304. PMLR.

Samvelyan, M.; Rashid, T.; Schroeder de Witt, C.; Farquhar, G.; Nardelli, N.; Rudner, T. G.; Hung, C.-M.; Torr, P. H.; Foerster, J.; and Whiteson, S. 2019. The StarCraft Multi-Agent Challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2186–2188.

Schaffernicht, E.; and Gross, H.-M. 2011. Weighted mutual information for feature selection. In *International Conference on Artificial Neural Networks*, 181–188. Springer.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *International Conference on Machine Learning (ICML)*.

Tan, M. 1997. Multi-Agent Reinforcement Learning: Independent versus Cooperative Agents. In *International Conference on Machine Learning*.

Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Xi Chen, O.; Duan, Y.; Schulman, J.; DeTurck, F.; and Abbeel, P. 2017. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30.

Wang, R.; He, X.; Yu, R.; Qiu, W.; An, B.; and Rabinovich, Z. 2020a. Learning efficient multi-agent communication: An information bottleneck approach. In *International Conference on Machine Learning*, 9908–9918. PMLR.

Wang, T.; Dong, H.; Lesser, V.; and Zhang, C. 2020b. ROMA: Multi-Agent Reinforcement Learning with Emergent Roles. In *International Conference on Machine Learning*, 9876–9886. PMLR.

Wang, T.; Wang, J.; Wu, Y.; and Zhang, C. 2019a. Influence-Based Multi-Agent Exploration. In *International Conference on Learning Representations*.

Wang, T.; Wang, J.; Zheng, C.; and Zhang, C. 2019b. Learning Nearly Decomposable Value Functions Via Communication Minimization. In *International Conference on Learning Representations*.

Yang, T.; Tang, H.; Bai, C.; Liu, J.; Hao, J.; Meng, Z.; and Liu, P. 2021. Exploration in deep reinforcement learning: a comprehensive survey. *arXiv preprint arXiv:2109.06668*.

Yu, C.; Velu, A.; Vinitsky, E.; Wang, Y.; Bayen, A.; and Wu, Y. 2021. The surprising effectiveness of mappo in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955*.

Zhang, K.; Yang, Z.; and Basar, T. 2019. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv preprint arXiv:1911.10635*.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Başar., T. 2018. Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents. In *International Conference on Machine Learning (ICML)*.

Zhang, T.; Rashidinejad, P.; Jiao, J.; Tian, Y.; Gonzalez, J. E.; and Russell, S. 2021a. Made: Exploration via maximizing deviation from explored regions. *Advances in Neural Information Processing Systems*, 34: 9663–9680.

Zhang, T.; Xu, H.; Wang, X.; Wu, Y.; Keutzer, K.; Gonzalez, J. E.; and Tian, Y. 2021b. Noveld: A simple yet effective exploration criterion. *Advances in Neural Information Processing Systems*, 34: 25217–25230.

Zheng, L.; Chen, J.; Wang, J.; He, J.; Hu, Y.; Chen, Y.; Fan, C.; Gao, Y.; and Zhang, C. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems*, 34: 3757–3769.