Social Motivation for Modelling Other Agents under Partial Observability in Decentralised Training

Dung Nguyen , Hung Le , Kien Do , Svetha Venkatesh , Truyen Tran
Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Geelong, Australia
{dung.nguyen,thai.le,k.do,svetha.venkatesh,truyen.tran}@deakin.edu.au

Abstract

Understanding other agents is a key challenge in constructing artificial social agents. Current works focus on centralised training, wherein agents are allowed to know all the information about others and the environmental state during training. In contrast, this work studies decentralised training, wherein agents must learn the model of other agents in order to cooperate with them under partially-observable conditions, even during training, i.e. learning agents are myopic. The intrinsic motivation for artificial agents is modelled on the concept of human social motivation that entices humans to meet and understand each other, especially when experiencing a utility loss. Our intrinsic motivation encourages agents to stay near each other to obtain better observations and construct a model of others. They do so when their model of other agents is poor, or the overall task performance is bad during the learning phase. This simple but effective method facilitates the processes of modelling others, resulting in the improvement of the performance in cooperative tasks significantly. Our experiments demonstrate that the socially-motivated agent can model others better and promote cooperation across different tasks.

1 Introduction

Humans understand others through active social interactions [Tomasello, 2009]. When we cooperate with others to achieve a goal and when experiencing a setback, we seek information about the ambient context and our partners [Swann *et al.*, 1981]. Thus modelling others is integral to our worldview.

Modelling other agents is challenging [Albrecht and Stone, 2018; Rabinowitz *et al.*, 2018], wherein the modelling process is often realised as an auxiliary task for each agent (1) to either construct a better representation of the observation [He *et al.*, 2016; Hernandez-Leal *et al.*, 2019; Zintgraf *et al.*, 2019; Papoudakis *et al.*, 2021; Ndousse *et al.*, 2021; Gu *et al.*, 2021] or (2) to generate predictions that are useful for the decision-making process (e.g. actions [Jaques *et al.*, 2019; Lowe *et al.*, 2017; Wen *et al.*, 2021], goals [Raileanu *et al.*, 2018], or individual or joint value functions [Chitnis *et al.*, 2019]). However, these methods are based on the centralised-training decentralised-execution (CTDE) paradigm wherein agents are trained under the assumption that they know all about the world's states and can access others' observations. A recent approach [Papoudakis *et al.*, 2021] has pioneered to model other agents using partial observations; however, this work still makes a strong assumption about using the other agents' observations during training. *Thus the problem of modelling other agents without access to their full observation is still open.*

In human research, the hypothesis that we tend to be more active in seeking information about others when losing control is supported by evidence in [Swann *et al.*, 1981]. Inspired by this, we address the aforementioned open problem by introducing a new social reward that encourages agents to meet each other so that they can do better modelling. Prior works have investigated social motivation (SM) to facilitate cooperative learning in multi-agent reinforcement learning [Khan *et al.*, 2018; Zheng *et al.*, 2021]. Influencing others was employed as intrinsic motivation (**IM**) in [Jaques *et al.*, 2019] to improve cooperation, and the method was extended to decentralised training by learning a model of other agents to generate a prediction of the others' next action. The work in [Jaques *et al.*, 2019] refers to staying nearby each other as a side-effect of encouraging them to model each other. In other words, encouraging agents to model others will *implicitly* bring agents near each other. However, in reality, the agents often ignore modelling other agents since it is a difficult auxiliary task.

In this work, we propose intrinsic motivation for artificial agents to encourage them to meet and understand others, especially when experiencing a utility loss. If the agents meet often, they can collect more relevant information about others, and this simple but effective method underpins our modelling. Furthermore, better understanding between agents significantly improves the performance in cooperative tasks.

The intrinsic reward, also called social motivation (**SM**), is proportional to the number of time steps in which agents are in each other's vicinity. This contributes to the final reward by being modulated through an *adaptive social motivation coefficient (**ASMC**).* This ASMC encourages agents to come close when the task performance reduces or if it evaluates that its model of the other agent is incorrect.

Our setting is decentralised training; that is, the agent can only access its own observations. We demonstrate our results on a) a partial-observable decentralised-training variant of the speaker-listener game [Mordatch and Abbeel, 2018; Papoudakis *et al.*, 2021], in which agents need to understand each other to communicate goals and achieve high team performance; and b) Facilitating cooperation in a level-based foraging environment [Christianos *et al.*, 2020]—a mixed cooperative-competitive setting, in which each agent can choose to be selfish or cooperative in a group. We show empirically that such socially motivated agents are effective in cooperative and mixed cooperative-competitive settings, outperforming agents with other types of intrinsic motivations.

Our contributions are a) Formulation of a novel, explicit social motivator for agents that is controlled adaptively by both the team performance and the efficacy of modelling the other agents; b) Demonstration that agents using our social motivation outperform agents that are not socially-motivated both in cooperative and mixed cooperative-competitive settings.

2 Related Works

2.1 Modelling Other Agents (MOA)

The work in [Papoudakis *et al.*, 2021] predicts the observation of others during the training phase; however, it is only a feasible assumption in the centralised-training paradigm. In [Ndousse *et al.*, 2021], authors augmented the agent with the reconstruction loss as an auxiliary task to construct the latent space that is useful for predicting the next state of the environment. This is similar to the local version in [Papoudakis *et al.*, 2021], and both, as we show in our experiments, do not behave well in the decentralised-training setting. Dealing with modelling under the partial observability problem, in [Gu *et al.*, 2021], the authors derived a mutual information loss between the policy representation and the teammates' situation to learn the encoder. However, it requires the states of the environment during training, while we do not provide any information except the observation of the agent itself, the so-called decentralised-training paradigm [Zhang *et al.*, 2018; Iqbal and Sha, 2019; Qu *et al.*, 2019; Tan, 1993; de Witt *et al.*, 2020]. We are the first work that to apply the bias in humans about social motivation [Pittman and Pittman, 1980; Swann *et al.*, 1981] to *explicitly* encourage agents to stay close to each other to facilitate learning the model of other agents under partial observability and decentralised-training scenario. Current research employs the model of others for different purposes: (1) as a condition to the decision-making (DM) process either by using features generated by the model of others or using the prediction such as action, intention, and goal, etc.—the neural networks that generate this information will not be trained by signals from the DM networks [Papoudakis *et al.*, 2021; Papoudakis and Albrecht, 2020; Hernandez-Leal *et al.*, 2019]; (2) as an auxiliary task to get a better representation of the observation (shared feature with the DM networks)—the model of others networks and the DM networks will share parameters at some levels [Hernandez-Leal *et al.*, 2019]; or (3) as a mechanism to produce more training signals during the learning phase—the model of others is used to shape rewards to learn social influence [Jaques *et al.*, 2019]. Our work is in line with the first type of model in the listed methods, which contains different modules for modelling other agents and for its policy.

2.2 Social Motivation for Cooperative RL

The phenomenon that a human is intrinsically motivated by novel situations is conveyed early in [White, 1959; Berlyne, 1966]. In artificial intelligence (AI), it is first known as computational curiosity [Schmidhuber, 1990; Meyer and Wilson, 1991], and is an emerging research in developmental robotics [Oudeyer and Kaplan, 2009]. In developing single reinforcement learning (RL) agents, intrinsic motivation is considered as an instrument to develop exploratory behaviour [Barto, 2013]. The first line is count-based intrinsic motivation, in which the novelty of a state or observation is evaluated by counting the number of time steps that the agents visit a particular state or receive the observation from the environment [Bellemare *et al.*, 2016; Tang *et al.*, 2017]. The second line is surprise-based intrinsic motivation [Pathak *et al.*, 2017]. This method motivates the agent to explore states and observations that are unpredictable. Although encouraging agents to meet others more can also be considered a form of curiosity about the world since it tends to reveal novel and interesting information about others—the world surrounding the agents, our proposed method and results also highlight the importance of focusing on other agents in social scenarios, as discussed in [Lerer and Peysakhovich, 2017; Eccles *et al.*, 2019; Hughes *et al.*, 2018].

Finding computational social intrinsic motivation is a fruitful direction [Khan *et al.*, 2018]. In MARL, research mainly focuses on utilising agents with surprise-based intrinsic motivation. The idea of intrinsic motivation from single RL is directly applied to MARL by learning the individual intrinsic reward for each agent to stimulate them to behave differently, leading to diversity amongst the group [Du *et al.*, 2019]. In [Chitnis *et al.*, 2019], each agent predicts the effects of joint actions and regularises each agent's action toward this joint action to achieve the collaborative task. Different to ours, this method only works in the centralised-training paradigm. Episodic memory is also employed to compute the social intrinsic motivation [Zheng *et al.*, 2021]. Prior works also proposed intrinsic motivation in MARL which is inspired by human social interaction. In [Jaques *et al.*, 2019], authors proposed to motivate agents to act in a way that can change the action of others. In this work, the model of others is to generate actions of others to make causal inferences in the independent training. Utilising the model of other agents to generate action to execute the causal reasoning will *implicitly* cause the agents to be close to each other, i.e. staying near others is considered to be a *side-effect* of the regulariser to predict others' actions. However, it is merely the case in reality since if the agents cannot soon find the benefit of learning from others, they will not focus on the modelling task. We, instead, argue that in the decentralised-training paradigm, it is crucial to *explicitly* encourage agents to be visible to each other. In other words, our agents are intrinsically motivated to model others. Being an orthogonal method, our reward can also help to increase the performance of agents proposed in

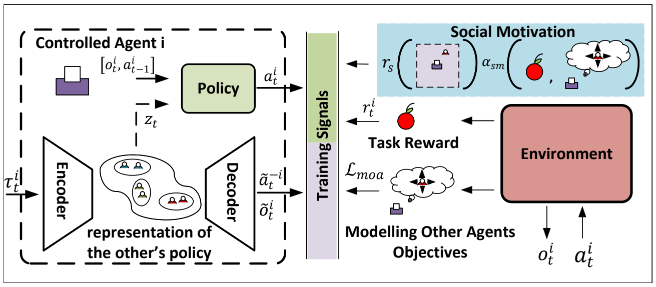


Figure 1: **Modelling other agents with social motivation.** During training, the controlled agent ( ) learns to represent its opponent's policy ( ) by reconstructing its observation and by predicting the other agent's actions based on its history τ_t^i. The controlled agent is motivated to meet other agents by the reward r_t^s that is proportional to the number of time steps it sees an agent in its field of view ( ). r_t^s is further regularised by the adaptive social motivation coefficient (α_t^{sm}) that increases if the task reward ( ) diminishes or its model of others ( ) is poor.

[Jaques *et al.*, 2019] and different count-based and surprise-based intrinsic motivations. In social robotics, this is known as *social drive* [Breazeal, 2004], where the robots have their own threshold for a social level to manage social interaction.

Other methods, such as [Wang *et al.*, 2019; Fayad and Ibrahim, 2021], proposed intrinsic motivation to act so that the controlled agent can encourage other agents to explore the states of the world. In these methods, the intrinsic reward was defined by the difference between the action value and the counterfactual function. Zhang et al. [Zhang *et al.*, 2021] proposed an intrinsic reward to guide agents' attention toward different regions and tasks to improve performance in dyadic collaborative manipulation. In [Yoo *et al.*, 2022], authors implemented the idea of influence-seeking behaviour in social agents by directly measuring the variance of the expectation of return. Agents can also be motivated to behave more predictably to their partners [Ma *et al.*, 2022]. The work in [Hussenot *et al.*, 2021] proposed intrinsic motivation to learn exploratory behaviours from the demonstrations of others. However, none of the listed methods considers *the motivation to learn about others*.

3 Preliminaries

3.1 Problem Formulation

We consider Multi-Agent Partially Observable Markov Decision Process (MA-POMDP) settings defined by the tuple $\langle \mathcal{S}, \mathcal{T}, \mathcal{A}_j, \mathcal{R}, \Omega_j, N \rangle$ with N is the number of agents and $j \in \{1 \dots N\}$ is the index of j^{th} agent in the environment. Here, $\mathcal{S}$ is the world state space, Ω_j is a function that map the world state space to the observation space $\mathcal{O}_j$ characterised by each agent $\Omega_j : \mathcal{S} \mapsto \mathcal{O}_j$, $\mathcal{T}$ is the transition function that returns the next state of the environment based on the current state and the joint action of all agents in the environment $\mathbf{a} \in \mathcal{A}^N = \times_{j \in \{1 \dots N\}} \mathcal{A}_j$, i.e. $\mathcal{T} : \mathcal{S} \times \mathcal{A}^N \mapsto \mathcal{S}$. At time step t, the agent j^{th} will observe the observation $o_t^j \in \mathcal{O}_j$ and take an action $a_t^j \in \mathcal{A}_j$. Each agent receives the reward $\mathcal{R}_j : \mathcal{S} \times \mathcal{A}^N \mapsto \mathbb{R}$ and tries to maximise the expected return $\mathbb{E}\big[\sum_{t=0}^{T} \gamma^t r_t^j\big]$, where T is the episode length and $\gamma \in (0, 1)$ is the discount factor. In this paper, we deal with the decentralised-training problem in which the agents can only access its observation o_t^j during training [Tan, 1993; de Witt *et al.*, 2020]. In our setting, we aim to find the optimal policy π^i of *a controlled agent* i^{th}, denoted as the superscript i, to work in a group with other agents, denoted as the superscript $-i$, sampled from a fixed set of L policies $\Pi^{-i} = \{\pi_l^{-i}\}_{l=1 \dots L}$. In case the controlled agent maintains a model of other agents, we also refer to other agents $(-i)$ as *modelled agents*. The solution to our problem is finding the optimal policy π_*^i of the controlled agent i over the fixed policy set of opponents Π^{-i}, i.e.

$$\pi_*^i = \text{argmax}_{\pi^i} \mathbb{E}_{\pi^{-i} \sim \Pi^{-i}} \left[\sum_{t=0}^{T} \gamma^t r_t^i \right].$$

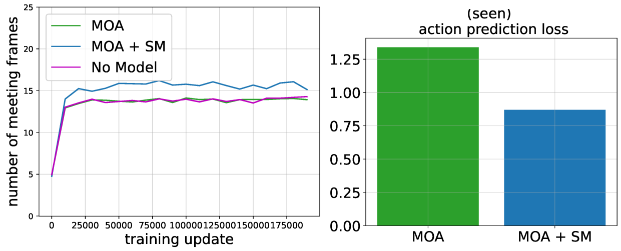


Figure 2: **The modelling-and-meeting relationship.** The average number of meetings in one episode during training (left) and the action prediction loss when the agent sees the others (right). MOA is agent with model of others but *not* socially motivated, MOA + SM is agent with model of others and socially motivated, and No MOA is agent *without* model of others.

3.2 Policy Conditioned on Other's Policy Representation

The general framework of the policy that is conditioned on the representation of other's policy is shown in Figure 1. The controlled agent i encodes its history $\tau_t^i = \{\langle o_{t'}^i, a_{t'}^i \rangle\}_{t'=1 \dots t-1}$, which *possibly* contains information about others, into a vector that represents the policy of the modelled agents by a neural network $z_t = \text{encoder}(\tau_t^i)$. To capture the temporal characteristic of the trajectory, we used the long-short term memory for implementing the encoder. This vector z_t is used as an input to the decoder for reconstructing the observations of the controlled agent i ($\tilde{o}_t^i$) and predicting the actions of modelled agents $-i$ ($\tilde{a}_t^{-i}$). These are two objectives to train the agents to have the ability to model other agents (included in $\mathcal{L}_{moa}$) that are specified in the Section 4.3. In the decentralised-training paradigm, agents will lack information about others due to the partial observation in the training phase. Hence, the implicit assumption to construct a good model about others is that the observation of the agent i contains the trajectory of other agents $-i$. The more times the agent i sees others, the more data and information it can obtain to build a sufficient model about others.

3.3 The Modelling-and-Meeting Relationship

This section analyses the relationship between modelling other agents and the number of meetings between agents, e.g. staying close and observing others. Figure 2 shows a partial result from agents trained in our experiment. In this figure, MOA stands for agents with the model of others but *not* socially motivated, MOA + SM denotes agents with the model of others and socially motivated, and No MOA denotes agents *without* the model of others. MOA does not meet others more than No MOA (purple and green curves are almost at the same level). However, explicitly encouraging agents to meet others, MOA + SM, increases the number of meeting other agents (see Fig. 2 (left), the blue curve is at higher level than both the purple and green curves) and improves the performance in modelling other agents (see Fig. 2 (right), the blue bar is lower than the green bar).

To improve the performance in cooperative tasks that requires an understanding of others, we need to address two failure cases for agents under partial observability: (1) **Case 1**: reward hacking while the agent only focuses on learning to reduce the loss in reconstructing the scene, e.g. going to a region without others. In this situation, because there are no other agents, it can enjoy the pleasure of predicting the static environment well (but not beneficial to the primary task); (2) **Case 2**: ignorance of modelling other agents. During the earlier stages, agents can learn the policy to optimise the portion of task rewards that do not require coordination.

4 Approach

4.1 Motivation to Understand Others

In the previous section, we argue that the challenge of modelling other agents under partial observation in decentralised training mainly comes from the fact that agents can not see each other during training. This causes the lack of training signal, which may block agents from learning a good model of others. Therefore, we propose to motivate the agent toward actions that can facilitate social interaction, e.g. meeting each other more during the training stage.

At every time step t, to motivate the controlled agent to meet others, we augment its reward with r_t^s. The reward r_t^s is proportional to the number of frames that the agent observes others in a set of experiences during a partial trajectory $\tau_{t':t'+T_b} = \left\{\left\langle o_t^i, a_t^i \right\rangle\right\}_{t=t',\dots,t'+T_b}$ where $T_b < T$ is the length of counting window. Specifically, if the controlled agent i has an experience $\left\langle o_t^i, a_t^i \right\rangle \in \tau_{t':t'+T_b}$, we first count the number of observations that the agent observes others

$$c_t^s = \sum_{\tau_{t':t'+T_b}} \mathbb{I}\left[-i \text{ in } o_t^i\right] = \sum_{\tau_{t':t'+T_b}} \mathbb{I}\left[\right.$$ $$\left.\right]$$

with $\mathbb{I}\left[\cdot\right]$ is an indicator function. We then compute the reward for stimulating agents to meet by

$$r_t^s = e^{K_s c_t^s}$$

where K_s is a coefficient that regulates how the augmented reward is sensitive to having other in the field of view.

4.2 Seeking Information

The controlled agent is *motivated to see others* if there is a decrease in its rewards. In other words, the agent *actively seeks information to improve its model of others* when the returns are reduced and focuses on refining the policy when the model about others is adequate. This allows the agent to simultaneously find the optimal policy while maintaining the motivation to understand other agents since the controlled agent can come back to model other agents at any time during the training process. To model this adaptive behaviour, we introduce an *adaptive social motivation coefficient* α_t^{sm} *(ASMC)*, which is updated every window of episodes. In a window of episodes w^{th}, we denote the minimum reward received as $r_w^{\min}$. The coefficient is defined as a non-linear function, which is formally written as

$$\alpha_t^{sm} = \underbrace{\frac{1}{2}\left(\tanh\left(-K_w \Delta_w\right) + 1\right)}_{\text{task performance}} - \underbrace{K_m \text{log} p\left(a_t^{-i} | \tau_t\right)}_{\text{ability to model other}}$$

where $\Delta_w = r_w^{\min} - r_{w-1}^{\min}$ is the difference between the minimum returns in two consecutive window of episodes, K_w indicates the sensitive of the decrease in returns to α_t^{sm}, and K_m weights the importance of modelling other agents.

The first term will *increase when the task performance reduces*. Intuitively, the second term encourages the agent to follow and model other agents *if it evaluates that the model of the other is incorrect*, which is computed by the accuracy of predicting other agents' actions. Furthermore, this adaptive social motivation coefficient α_t^{sm} is employed to regularise the effect of the reward for encouraging agents to meet r_t^s, which is detailed in the next section.

4.3 The Social Motivation Reward

The social reward based on social motivation is constructed as

$$r_t^{sm} = \alpha_t^{sm} r_t^s.$$

Because both α_t^{sm} and r_t^s are positive, it implies that our approach only encourages the controlled agent to meet others. We balance between understanding others and achieving high task performance. This social reward differs from other intrinsic social rewards proposed in the literature: it focuses on improving the model of other agents under partial observation in decentralised training.

The agents will optimise the reward

$$r_t = r_t^i + r_t^{sm} = r_t^i + \alpha_t^{sm} r_t^s,$$

which includes: the primary task reward (r_t^i) and the proposed social motivation reward (r_t^{sm}). During the training process, the modelling other agent process is encouraged by minimising the loss $\mathcal{L}_{moa} = \mathcal{L}_{obs} + \mathcal{L}_{act}$ with the observation reconstruction loss and the action prediction loss are:

$$\mathcal{L}_{obs} = \left(o_t^i - \tilde{o}_t^i\right)^2, \text{and}$$

$$\mathcal{L}_{act} = -\text{log} p\left(a_t^{-i} | \tau_t\right),$$

respectively.

5 Experiments

We conduct a suite of experiments to show:

- The effect of our method on the performance of agents with/without the model of others, hence, proving the improvement is via better modelling of other agents;
- That our proposed approach outperforms other intrinsic motivations in the cooperative settings; and,
- The analysis of the effect of social motivation on cooperation in the mixed cooperative-competitive setting.

5.1 Cooperative Setting

Setting: The Speaker-Listener Game

The speaker-listener game is a cooperative game in which two agents need to communicate to achieve goals. This game has two agents—one is the controlled agent, the other is the modelled agent—and three landmarks in different colours. Based on the world observation and the message received from its teammate, the agent will take physical actions to navigate to the landmark (goal) that has the same colour as the agent. Since each agent cannot observe its own colour, it relies on the message sent by its teammate. Because the game is cooperative, the reward that the controlled agent tries to optimise is the team reward, i.e. the average of the negative distances between team members and their goals. Each episode is terminated after 25 timesteps or 50 timesteps. In the original implementation of the game and a recent investigation in modelling other agents [Mordatch and Abbeel, 2018; Papoudakis *et al.*, 2021], both agents can share observations during training. We study a variant with the changes in the observability of agents during both training and execution. In our setting, agents are in partial observability in both training and execution, which means they can only observe others' positions if others are in their field of view.

Training Other Agents

Before joining the training process, the agent navigation policy of $-i$ is pre-trained with a fixed communication policy. We trained 10 agents by MADDPG [Lowe *et al.*, 2017] until they converged to the optimal navigation behaviour. We conducted 5 self-play games to train 10 agents, i.e. each game has two agents with the same architecture but different initialisation. The reward is given as the original speaker-listener games described in [Mordatch and Abbeel, 2018]. This stage of pre-training a set of fixed policies is well-known for research in modelling other agents [Papoudakis *et al.*, 2021].

Baselines

In figures onward, the socially-motivated agent with the adaptive social motivation coefficient (ASMC) α_t^{sm} is denoted as SM + ASMC. The socially-motivated agent *without* the ASMC is denoted as SM and has fixed $\alpha^{sm} = 0.001$. We optimised the controlled agent's policy using A2C algorithm and set hyper-parameters $K_s = 5.0, K_w = 15.0, K_m = 0.1$.

We compare the performances of our agents which have the model of others and are socially motivated (SM and SM + ASMC) to (1) agents *without* model of other agents (No MOA), (2) agents that are socially motivated but *without* model of other agents (No MOA + SM), and (3) agents with model of other agents but are *not* socially motivated (LIAM-Local) [Papoudakis *et al.*, 2021]. We also considered different intrinsic motivations, including count-based IM (IC), surprised-based IM (IS), and causal influence (CI) agents.

Motivation to Model Other Agents Helps Improve Team Performance in Cooperative Tasks

Figure 3 (left column) shows team rewards, i.e. the performance of learning agents in the speaker-listener task. To analyse the effects of social motivation on the performance, we decompose the team reward into the reward to the modelled agent r^{modelled} (the negative distance between the controlled agent and its goal) in Fig. 3 (middle column), and the reward to the controlled agent $r^{\text{controlled}}$ (between the modelled agent and its goal) in Fig. 3 (right column). *First*, agents without the model of others perform worse (black and purple curves) because this task requires understanding others' interpretation of messages. *Second*, the agent augmented with the model of others behaves better than the agent without the model of others (the green curve); however, it is far worse than agents with models that are encouraged to meet others (the blue and red curves). The improvement mainly comes from the performance of the modelled agent, i.e. the controlled agent sends more precise messages that match with the understanding of the modelled agent. *Third*, our SM can only help in case the agent has the model of other agents. Otherwise, it could not increase the performance of the team, i.e. our new social reward is to stimulate agents to understand each other. *Finally*, agents with ASMC achieve higher team returns than agents with fixed α^{sm}. Figure 4 shows that our agents can predict the other's actions significantly better than the agent without social motivation. Results as shown in Fig. 4 and 3 (right column) empirically prove that the improvement in the team reward and the reward received by our agent is due to the ability to model others better.

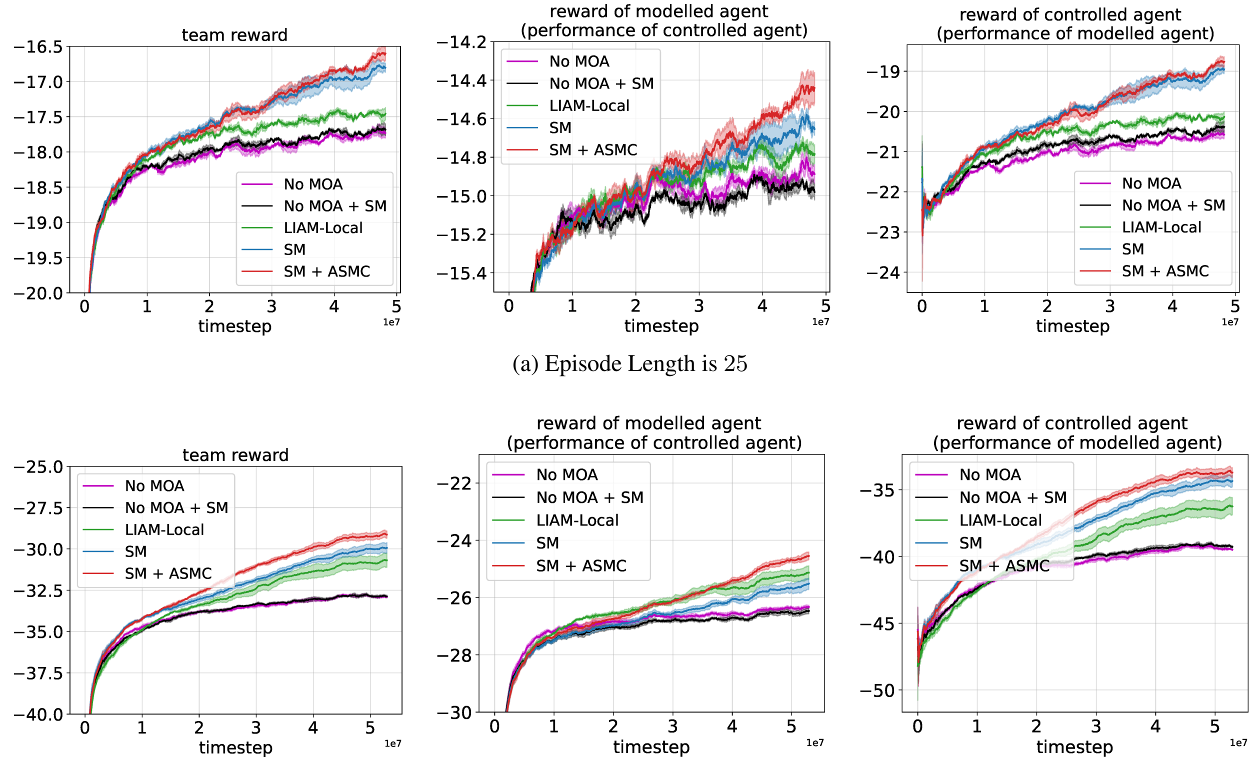
Comparison with Other Intrinsic Motivations (IMs)

Encouraging agents to meet reduces the reconstruction loss compared to the cases where agents are motivated by other IMs such as novelty-based, surprise-based, and influence-based intrinsic motivation. This is shown in Table 1.

5.2 Mixed Cooperative-Competitive Setting

Setting: The Level-based Foraging

In this setting, two agents try to collect four apples in the grid world. Agents can only partially observe the world. Each agent has its level, as does each apple. They can pick an apple if the sum of all agents' levels is higher than the apple level. If there is an apple nearby at the lower level, one can choose to collect this apple instead of cooperating with others individually. Therefore, this setting has mixed cooperative-competitive properties, i.e. one can selfishly collect the apple with a lower level and cooperate with others only to collect the high-level apples. Understanding whether the opponent is cooperative or competitive is critical to achieving high performance in this environment. One episode is terminated when all apples are collected or the episode length (50 time steps) is reached.



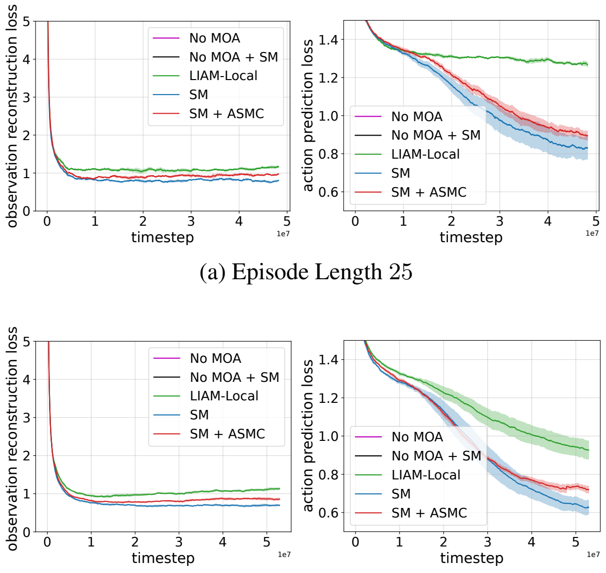
(a) Episode Length is 25

(b) Episode Length is 50

Figure 3: **Performance in decentralised-training cooperative setting (higher is better).** The team return (left column), the return received by the modelled agent (middle column), and the return received by the controlled agent (right column) vs. the training timestep when trained with the episode length $T = 25$ and $T = 50$. Social motivation reward significantly helps to improve the performance of the agent in a cooperative task, especially when the episode lasts long and there is enough time for agents to regulate between acting to model and acting towards achieving a high reward in the task. With the adaptive social motivation coefficient, the agent can balance between learning to achieve its own task (higher performance in the middle figure) and learning to model other agents, leading to obtaining higher team rewards. Encouraging agents to meet others does not help to increase the performance if agents do not have models about others.

	Team Reward (↑)	r^{modelled} (↑)	$r^{\text{controlled}}$ (↑)	Act. Prediction Loss (↓)
No MOA	-17.724 ± 0.051	-14.841 ± 0.039	-20.608 ± 0.074	
No MOA + SM	-17.660 ± 0.052	-14.905 ± 0.035	-20.413 ± 0.076	
LIAM-Local	-17.385 ± 0.068	-14.757 ± 0.060	-20.013 ± 0.080	1.224 ± 0.031
Intrinsic Count (IC)	-17.263 ± 0.117	-14.649 ± 0.054	-19.877 ± 0.180	1.164 ± 0.052
Intrinsic Surprise (IS)	-17.262 ± 0.079	-14.701 ± 0.040	-19.824 ± 0.140	1.206 ± 0.039
Causal Inference (CI)	-17.181 ± 0.065	-14.649 ± 0.055	-19.738 ± 0.102	1.118 ± 0.042
Social Motivation (SM)	-16.822 ± 0.094	-14.619 ± 0.069	-19.026 ± 0.121	$\mathbf{0.844 \pm 0.048}$
SM + ASMC	$\mathbf{-16.636 \pm 0.036}$	$\mathbf{-14.487 \pm 0.026}$	$\mathbf{-18.785 \pm 0.050}$	0.861 ± 0.033
IC + SM	-16.563 ± 0.142	-14.438 ± 0.088	-18.688 ± 0.196	$\underline{\mathbf{0.759 \pm 0.031}}$
IS + SM	-16.671 ± 0.075	-14.543 ± 0.042	-18.800 ± 0.126	0.763 ± 0.030
CI + SM	$\underline{\mathbf{-16.531 \pm 0.072}}$	$\underline{\mathbf{-14.424 \pm 0.046}}$	$\underline{\mathbf{-18.639 \pm 0.102}}$	0.786 ± 0.027
IC + SM + ASMC	-16.782 ± 0.068	-14.518 ± 0.051	-19.046 ± 0.100	0.892 ± 0.034
IS + SM + ASMC	-16.626 ± 0.046	-14.444 ± 0.049	-18.808 ± 0.072	0.846 ± 0.022
CI + SM + ASMC	-16.709 ± 0.124	-14.528 ± 0.063	-18.891 ± 0.189	0.852 ± 0.028

Table 1: **Quantitative comparison on cooperative task.** $r^{\text{controlled}}$ shows the performance of the modelled agent, which indicates how well our agent can model the other to help the other to achieve the task. r^{modelled}, instead, shows the performance of the controlled agent on achieving its own task. The r^{modelled} does not different across all, but the $r^{\text{controlled}}$ are different. This illustrates the improvement from the team reward actually comes from the improvement in modelling other agents. We also could observe the lower action prediction loss while social motivation is applied.



(a) Episode Length 25

(b) Episode Length 50

Figure 4: **Losses to encourage modelling other agents during training (lower is better).** The socially-motivated agent can learn to better reconstruct observations and predict the actions of others.

Training Other Agents

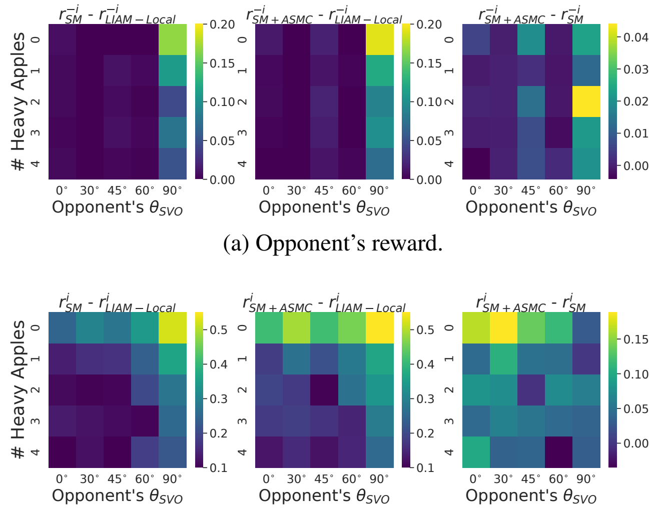
Our population of fixed policy agents are diverse in the *social value orientations* (SVO) [Murphy and Ackermann, 2014; Griesinger and Livingston Jr, 1973; Liebrand and McClintock, 1988]—a measure in social psychology to indicate a preference of weighting rewards between themselves and others. In AI, this concept was applied to estimate human social behaviour in driving to build better autonomous vehicles [Schwarting *et al.*, 2019] or to create distinct populations of artificial social learning agents in [McKee *et al.*, 2020]. We pre-trained 5 pairs of agents with different SVOs $\theta_{SVO} \in \{0°, 30°, 45°, 60°, 90°\}$. For example, $\theta_{SVO} = 0°$, the agent is individualistic, while agent with $\theta_{SVO} = 90°$ is altruistic.

Evaluation

In this setting, to analyse the effect of the social motivation and the adaptive social motivation coefficient, we paired up trained agents with different opponents that have different SVOs. Each pair of agents is evaluated in environments in which there are different numbers of *heavy apples*—apples that have a higher level than all agents and require agents to cooperate in collecting successfully. The difference in the number of heavy apples will create a different demand for cooperation to achieve in this social environment. If there are no heavy apples, agents do not need to cooperate. However, if all apples are heavy, they need to cooperate in collecting.

Results

In Figure 5, we compare the difference between pairs of methods (left column) SM and LIAM-Local, (middle column) SM+ASMC and LIAM-Local, and (right column) SM+ASMC and SM, for increasing levels of θ_{SVO}, taking agents from individualistic to fully altruistic in different environment condition. *First*, we observe that having SM will improve the performance of agents that have the model of others. This is shown by the fact that there are no negative values in the left and middle columns in Fig. 5(a,b). *Second,* when paired up with individualist agents (low SVOs), the difference between the performance of the socially-motivated agent and the agent without SM is higher when there are fewer heavy apples. It is because in an environment where every apple can be collected, if our agent realises that the other agent is individualist, it will try to collect as much as possible; hence modelling the opponent in earlier steps of the episode improves the performance of the controlled agent. In contrast, if all apples in the environment are heavy, then our agent cannot collect more apples because of the opponent's uncooperative behaviour. *Third*, interestingly, agents motivated to meet others during training improves the performance of their partner who has high SVO, i.e. who is altruistic and willing to cooperate, as shown in Fig. 5(a), leading to higher performance of the team. *Finally*, the ASMC helps to increase the reward of our agents when it is matched with more selfish agents in environments that do not require cooperation.



(a) Opponent's reward.

(b) Controlled agent's reward.

Figure 5: **Pairwise comparison of performance between methods in mixed cooperative-competitive setting for (a) the opponent and (b) the controlled agent, with different levels of altruism.** Comparison between (left column) SM and LIAM-Local, (middle column) SM+ASMC and LIAM-Local, and (right column) SM+ASMC and SM. Each cell shows the difference between the return obtained by agents in environments with differing numbers of heavy apples. The lighter colour means a higher difference.

6 Conclusions

We propose a novel social motivation for artificial agents under partial observation in decentralised training. Different from other intrinsic motivations, we explicitly motivate agents to meet others to acquire information to build a better model about others, especially when the agent experiences a decrease in its performance or if the model of others is poor during training. We empirically show that this social motivation increases the performance of agents in cooperative tasks.

Acknowledgements

This research was partially funded by the Australian Government through the Australian Research Council (ARC). Prof. Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

References

[Albrecht and Stone, 2018] Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[Barto, 2013] Andrew G Barto. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*, pages 17–47. Springer, 2013.

[Bellemare *et al.*, 2016] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *NIPS*, 29, 2016.

[Berlyne, 1966] Daniel E Berlyne. Curiosity and exploration: Animals spend much of their time seeking stimuli whose significance raises problems for psychology. *Science*, 153(3731):25–33, 1966.

[Breazeal, 2004] Cynthia Breazeal. *Designing sociable robots*. MIT press, 2004.

[Chitnis *et al.*, 2019] Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. Intrinsic motivation for encouraging synergistic behavior. In *ICLR*, 2019.

[Christianos *et al.*, 2020] Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. In *NeurIPS*, 2020.

[de Witt *et al.*, 2020] Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the Starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

[Du *et al.*, 2019] Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. *NeurIPS*, 32, 2019.

[Eccles *et al.*, 2019] Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z Leibo. Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*, 2019.

[Fayad and Ibrahim, 2021] Ammar Fayad and Majd Ibrahim. Influence-based reinforcement learning for intrinsically-motivated agents. *arXiv preprint arXiv:2108.12581*, 2021.

[Griesinger and Livingston Jr, 1973] Donald W Griesinger and James W Livingston Jr. Toward a model of interpersonal motivation in experimental games. *Behavioral science*, 18(3):173–188, 1973.

[Gu *et al.*, 2021] Pengjie Gu, Mengchen Zhao, Jianye Hao, and Bo An. Online ad hoc teamwork under partial observability. In *ICLR*, 2021.

[He *et al.*, 2016] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *ICML*, pages 1804–1813. PMLR, 2016.

[Hernandez-Leal *et al.*, 2019] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. Agent modeling as auxiliary task for deep reinforcement learning. In *Proceedings of the AAAI conference on AI and interactive digital entertainment*, volume 15, pages 31–37, 2019.

[Hughes *et al.*, 2018] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *NeurIPS*, 31, 2018.

[Hussenot *et al.*, 2021] Léonard Hussenot, Robert Dadashi, Matthieu Geist, and Olivier Pietquin. Show me the way: Intrinsic motivation from demonstrations. In *AAMAS*, pages 620–628, 2021.

[Iqbal and Sha, 2019] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *ICML*, pages 2961–2970. PMLR, 2019.

[Jaques *et al.*, 2019] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *ICML*, pages 3040–3049. PMLR, 2019.

[Khan *et al.*, 2018] Md Mohiuddin Khan, Kathryn Kasmarik, and Michael Barlow. Toward computational motivation for multi-agent systems and swarms. *Frontiers in Robotics and AI*, 5:134, 2018.

[Lerer and Peysakhovich, 2017] Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

[Liebrand and McClintock, 1988] Wim BG Liebrand and Charles G McClintock. The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation. *European journal of personality*, 2(3):217–230, 1988.

[Lowe *et al.*, 2017] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *NIPS*, 30, 2017.

[Ma *et al.*, 2022] Zixian Ma, Rose E Wang, Li Fei-Fei, Michael S Bernstein, and Ranjay Krishna. Elign: Expectation alignment as a multi-agent intrinsic reward. In *NeurIPS*, 2022.

[McKee *et al.*, 2020] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duèñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. In *AAMAS*, pages 869–877, 2020.

[Meyer and Wilson, 1991] Jean-Arcady Meyer and Stewart W Wilson. *A possibility for implementing curiosity and boredom in model-building neural controllers*. MIT Press, 1991.

[Mordatch and Abbeel, 2018] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *AAAI*, volume 32, 2018.

[Murphy and Ackermann, 2014] Ryan O Murphy and Kurt A Ackermann. Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1):13–41, 2014.

[Ndousse *et al.*, 2021] Kamal K Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. Emergent social learning via multi-agent reinforcement learning. In *ICML*, pages 7991–8004. PMLR, 2021.

[Oudeyer and Kaplan, 2009] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers in neurorobotics*, page 6, 2009.

[Papoudakis and Albrecht, 2020] Georgios Papoudakis and Stefano V Albrecht. Variational autoencoders for opponent modeling in multi-agent systems. *arXiv preprint arXiv:2001.10829*, 2020.

[Papoudakis *et al.*, 2021] Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. Agent modelling under partial observability for deep reinforcement learning. *NeurIPS*, 34:19210–19222, 2021.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, pages 2778–2787. PMLR, 2017.

[Pittman and Pittman, 1980] Thane S Pittman and Nancy L Pittman. Deprivation of control and the attribution process. *Journal of Personality and Social Psychology*, 39(3):377, 1980.

[Qu *et al.*, 2019] Chao Qu, Shie Mannor, Huan Xu, Yuan Qi, Le Song, and Junwu Xiong. Value propagation for decentralized networked deep multi-agent reinforcement learning. *NeurIPS*, 32, 2019.

[Rabinowitz *et al.*, 2018] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *ICML*, pages 4218–4227. PMLR, 2018.

[Raileanu *et al.*, 2018] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In *ICML*, pages 4257–4266. PMLR, 2018.

[Schmidhuber, 1990] Jürgen Schmidhuber. Making the world differentiable: On using fully recurrent self-supervised neural networks for dynamic reinforcement learning and planning in non-stationary environments. *Institut für Informatik, Technische Universität München. Technical Report FKI-126*, 90, 1990.

[Schwarting *et al.*, 2019] Wilko Schwarting, Alyssa Pierson, Javier Alonso-Mora, Sertac Karaman, and Daniela Rus. Social behavior for autonomous vehicles. *PNAS*, 116(50):24972–24978, 2019.

[Swann *et al.*, 1981] William B Swann, Blair Stephenson, and Thane S Pittman. Curiosity and control: On the determinants of the search for social knowledge. *Journal of Personality and Social Psychology*, 40(4):635, 1981.

[Tan, 1993] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *ICML*, pages 330–337, 1993.

[Tang *et al.*, 2017] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *NIPS*, 30, 2017.

[Tomasello, 2009] Michael Tomasello. *Why we cooperate*. MIT press, 2009.

[Wang *et al.*, 2019] Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. In *ICLR*, 2019.

[Wen *et al.*, 2021] Ying Wen, Yaodong Yang, and Jun Wang. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. In *IJCAI*, pages 414–421, 2021.

[White, 1959] Robert W White. Motivation reconsidered: the concept of competence. *Psychological review*, 66(5):297, 1959.

[Yoo *et al.*, 2022] Byunghyun Yoo, Devarani Devi Ningombam, Sungwon Yi, Hyun Woo Kim, Euisok Chung, Ran Han, and Hwa Jeon Song. A novel and efficient influence-seeking exploration in deep multiagent reinforcement learning. *IEEE Access*, 10:47741–47753, 2022.

[Zhang *et al.*, 2018] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *ICML*, pages 5872–5881. PMLR, 2018.

[Zhang *et al.*, 2021] Minghao Zhang, Pingcheng Jian, Yi Wu, Huazhe Xu, and Xiaolong Wang. Dair: Disentangled attention intrinsic regularization for safe and efficient bimanual manipulation. *arXiv preprint arXiv:2106.05907*, 2021.

[Zheng *et al.*, 2021] Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *NeurIPS*, 34:3757–3769, 2021.

[Zintgraf *et al.*, 2019] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep reinforcement learning via meta-learning. In *ICLR*, 2019.