Goal Consistency: An Effective Multi-Agent Cooperative Method for Multistage Tasks

Xinning Chen[1*], **Xuan Liu**[1*†], **Shigeng Zhang**[2], **Bo Ding**[3] and **Kenli Li**[1]
[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]School of Computer Science and Engineering, Central South University, China
[3]School of Computer Science, National University of Defense Technology, China
{chenxinning,xuan_liu,lkl}@hnu.edu.cn, sgzhang@csu.edu.cn, bding@msn.com

Abstract

Although multistage tasks involving multiple sequential goals are common in real-world applications, they are not fully studied in multi-agent reinforcement learning (MARL). To accomplish a multi-stage task, agents have to achieve cooperation on different subtasks. Exploring the collaborative patterns of different subtasks and the sequence of completing the subtasks leads to an explosion in the search space, which poses great challenges to policy learning. Existing works designed for single-stage tasks where agents learn to cooperate only once usually suffer from low sample efficiency in multi-stage tasks as agents explore aimlessly. Inspired by human's improving cooperation through goal consistency, we propose **M**ulti-**A**gent **G**oal Cons**I**sten**C**y (MAGIC) framework to improve sample efficiency for learning in multi-stage tasks. MAGIC adopts a goal-oriented actor-critic model to learn both local and global views of goal cognition, which helps agents understand the task at the goal level so that they can conduct targeted exploration accordingly. Moreover, to improve exploration efficiency, MAGIC employs two-level goal consistency training to drive agents to formulate a consistent goal cognition. Experimental results show that MAGIC significantly improves sample efficiency and facilitates cooperation among agents compared with state-of-art MARL algorithms in several challenging multistage tasks.

1 Introduction

Multi-agent reinforcement learning (MARL) has shown superiority in cooperative multi-agent decision control problems, such as multi-player games, autonomous vehicles and traffic control [Gronauer and Diepold, 2021]. However, learning to cooperate becomes extremely difficult due to the combinatorial possibilities of agent interactions and the huge state space in increasingly complex multi-agent tasks, which may be even worse when the tasks contain multiple stages. Multistage tasks are common in real-word application, such as freight transportation tasks that require the cargo to pass through multiple delivery points in sequence according to the prescribed routes. Such multi-stage tasks require agents to cooperate to achieve multiple goals in sequence together. Exploring the collaborative patterns of different subtasks and the completion sequence of the subtasks leads to an explosion in the search space, which poses greater challenges to policy learning than in single-stage tasks.

To promote cooperation in large state spaces, existing works mainly focus on optimizing a joint goal via sharing and enhancing information at the agent level. [Yang *et al.*, 2018] reduces the input dimension by approximating the interactions within the population of agents with an average effect. NCC-Q [Mao *et al.*, 2020] learns cognitive consistency of neighborhood observations to encourage collaboration. Attention mechanisms are widely used to simplify the learning process by discovering the relationship among agents [Iqbal and Sha, 2019; Mao *et al.*, 2019; Ryu *et al.*, 2020; Liu *et al.*, 2020]. However, these algorithms struggle to learn optimal policies with aimless exploration in more complex multistage tasks with multiple goals.

We observe that achieving goal consistency is a core factor of successful teamwork in human society. For cooperative tasks with multiple goals in the real world, people tend to reach a consensus on which goal to achieve first before taking action. Obviously, goal consistency helps avoid useless exploration and improves cooperation efficiency. In other words, when the goals pursued are inconsistent, human individuals are unlikely to form coordinated behaviors to complete tasks. Take wildlife rescue as an example. In this task, a group of rescuers needs to rescue several wild animals together based on the severity of their injuries. If rescuers do not reach a consensus on the assistance goals, the conflict will hinder their cooperation, leading to failure in rescuing any animals timely. In contrast, a team with a consistent goal can quickly determine the assistance goal and achieve precise rescue.

Inspired by the above observation, we introduce goal consistency into multi-agent reinforcement learning and propose *Multi-Agent Goal Consistency* (MAGIC) framework to promote cooperation among agents in multistage tasks with multiple goals. We explicitly incorporate goal cognition into the

[*] These authors contributed equally to this work.

[†] Corresponding author

decision-making process by designing a goal-oriented actor-critic model that learns to cognitive goals and dynamically focuses on different goals in different stages. Moreover, taking advantage of centralized training, we obtain goal cognition from local and global views, which further motivates the two-level goal consistency training. By keeping consistent with each other through self-supervised learning and aligning individual goal with that of other agents, each agent forms a consistent goal cognition in local view and global view to improve sample efficiency and conduct effective exploration, thus achieving efficient cooperation.

We evaluate our methods on several challenging multistage tasks. Experiment results show that agents exhibit mutually cooperative behaviors with consistent goal cognition. Moreover, MAGIC significantly outperforms state-of-the-art MARL approaches in convergence speed and final cooperation performance.

2 Related Work

Among existing MARL algorithms, sharing and collecting information about other agents is a common way to achieve consensus for better cooperation. Learning to share information by communication has been exploited to accelerate cooperation [Sukhbaatar *et al.*, 2016; Jiang and Lu, 2018; Ding *et al.*, 2020; Lin *et al.*, 2021]. Under the paradigm of centralized training and decentralized execution (CTDE), MADDPG [Lowe *et al.*, 2017] has extended DDPG [Lillicrap *et al.*, 2016] to multi-agent settings by augmenting each agent with information about others in centralized critics. NCC-Q [Mao *et al.*, 2020] imposes perceptual consensus constraints on the agent and other agents within its fixed neighborhood to encourage collaboration. Mean Filed [Yang *et al.*, 2018] approximates the interaction within neighboring agents using an average effect to enable coordination between large-scale agents. Recent works explore the effectiveness of employing attention networks [Xu *et al.*, 2015] to process the information shared by other agents. Multiple-actor-attention-critic (MAAC) [Iqbal and Sha, 2019] introduces multi-head soft-attention into the centralized critics to focus on relevant agents' information. Multi-agent game abstraction algorithm [Liu *et al.*, 2020] combines soft-attention and hard-attention mechanisms to learn the relationship between agents for effective cooperation. However, these methods tend to accomplish a joint goal via sharing and enhancing information at the agent level, which generally find difficulty to learn effective cooperation in large state space with aimless exploration.

To improve sample efficiency, several attempts have been made to promote efficient exploration in MARL domain. IGASIL [Hao *et al.*, 2019] guide agents to imitate from the past good experiences and do more exploration around these high-reward regions. Rochico [Li *et al.*, 2021b] improves exploration efficiency based on reinforced organization control and hierarchical consensus learning. CM3 [Yang *et al.*, 2020] constructs multi-stage curriculum learning to address the difficulty of multi-agent exploration while focusing on single-stage tasks with multiple parallel goals. CMAE [Liu *et al.*, 2021] trains exploration policies to reach a shared goal state selected from restricted spaces. Some approaches [Mahajan *et al.*, 2019; Mahajan *et al.*, 2019; Li *et al.*, 2021a] have been proposed recently to encourage extensive exploration.

In this paper, we focus on multistage tasks with sequential goals. By considering consistency at the goal level to guide efficient exploration, our method improves sample efficiency and promotes multi-agent cooperation in multistage tasks.

3 Preliminaries

In this section, we formalize the multistage MARL problem as a multistage partially observable Markov game. A multi-agent learning problem that contains L stages is defined as a tuple $\langle S, \mathbf{A}, \mathbf{O}, \mathbf{R}, \mathbf{G}, Z, P, \gamma \rangle$, where S is the set of environment state s, and $\mathbf{G}$ is the set of goals to achieve. At each stage $l \in [1 \cdots L]$, agents learn to finish sub-goals $g_l \in \mathbf{G}$. $\mathbf{O} = [O_1, \cdots, O_N]$ is a finite set of joint observations. Each agent $i \in [1 \cdots N]$ observes $o_i \in O_i$ drawn from the observation function $Z : S \times \mathbf{A} \rightarrow \mathbf{O}$. Then, agent $i \in \{1 \cdots N\}$ chooses its own action $a_i \in A_i$, forming a joint action $\boldsymbol{a} \in \mathbf{A}$, which induces state transition according to the state transition function $P(s'|s, \mathbf{a}) : S \times \mathbf{A} \times S \rightarrow [0, 1]$. After that, each agent i receives a reward $r_i \in R_i : S \times \mathbf{A} \times \mathbf{G} \rightarrow \mathbb{R}$. Without loss of generality, all joint quantities over agents are denoted in bold. The objective for each agent is to learn a policy $\pi_i(a_i|o_i, g_l) : O_i \times A_i \times \mathbf{G} \rightarrow [0, 1]$ to maximize the expected cumulative reward received $\mathrm{E}[\sum_{t=0}^{H} \gamma^t r_i^t]$ over horizon H, in which $\gamma \in [0, 1)$ is a discount factor.

4 Methodology

In this section, we specifically introduce our multi-agent goal consistency (MAGIC) framework to facilitate cooperation in multistage tasks. First, inspired by the human decision-making process from goal cognition to action, we explicitly incorporate goal cognition into the learning process of agents, and design a goal-orientated actor-critic model, which learns two views of goal cognition. Second, we formally introduce two-level goal consistency training to drive agents to have consistent goal cognition, thus avoiding useless exploration and promoting cooperation.

4.1 Goal-oriented Actor-Critic Model

In complex multistage environments, it is difficult to learn useful strategies through aimless exploration. To motivate effective exploration, we propose a goal-orientated actor-critic model, by which agents learn to act and criticize based on the learned goal cognition. Moreover, taking advantage of centralized training and decentralized execution, the decentralized policies (actors) and the centralized critics learn the local and global views of goal cognition, respectively.

Goal Cognition in Local View

To derive goal-oriented policies for effective exploration, we propose a general goal cognition module (GCM) to learn the local view of goal cognition adaptively from local observations. For agent i, the observation o^i can be expressed as $o^i = \left\{o^i_{-g}, o^i_{g_1}, \cdots, o^i_{g_T}\right\}$, where $o^i_{g_j}$ is the information about the goal g_j observed by agent i, called goal-specific

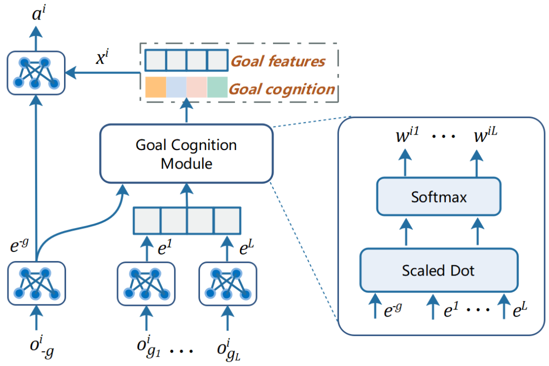


Figure 1: The goal-oriented policy network architecture of agent i.

observation. o^i_{-g} is the rest of o^i. Since o^i_{-g} contains all information about the task and environment except for the goal entities, we exploit o^i_{-g} to cognitive goals in decentralized policy networks. As shown in Fig.1, we first encode the local observation to get feature representations by embedding functions, where o^i_{-g} and $o^i_{g_j}$ are encoded into feature vector e_{-g} and e^j respectively. Then, the feature representation vectors are passed into the goal cognition module based on the soft-attention mechanism to recognize which goal is more important for the current stage. Specifically, the goal cognition module compares the embedding e^{-g} with e^j and passes the matching value between these two embeddings into a softmax function:

$$w^{i,j} = \frac{\exp(e^{-g}e^j)}{\sum_{j=1}^{L}\exp(e^{-g}e^j)}. \tag{1}$$

where $w^{i,j}$ denotes the importance of goal g_j for agent i.

The importance distribution of the goal-specific observations W^i_{local} exactly represents goal cognition of agent i in local view. A uniform distribution means that the agent has no clear cognition of the current goal, and it may act with equal consideration of all goals. The goal g_j with the greatest importance is seen as the agent's *selected goal*. Through goal cognition, agents pay more attention to the information of the selected goal, and take actions based on the feature vector e^{-g} and the weighted sum of each goal-specific observation:

$$a^i = \pi^i(o^i;\theta^i) = f_i(e^{-g}, x^i), \tag{2}$$

$$x^i = \sum_{j=1}^{L} w^{i,j}e^j, \tag{3}$$

where f_i is a fully connected layer. In this way, agents learn to dynamically focus on the observable goal information in each stage and act based on their local view of goal cognition.

Novelty of GCM. GCM empowers agents with the ability to explicitly cognitive stage goals by adopting a simple attention kernel. Unlike existing algorithms that learn the relationship between agents with attention mechanisms, GCM focuses on the importance of each goal in the current state, which is conducive to discovering the goal sequence of multistage tasks and facilitating targeted exploration.

Goal Cognition in Global View

During the centralized training, utilizing the information of all agents, the critics learn the global view of goal cognition and evaluate the value of observation-action pairs.

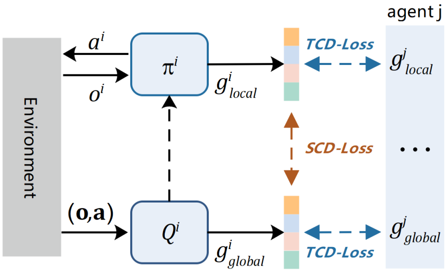


Figure 2: Training agents with two-level goal consistency by using the self cognitive dissonance loss (TCD-loss) and the team cognitive dissonance loss (SCD-loss).

Similar to the architecture of policy networks, the critic Q^i for agent i derives goal-specific observation from o^i. Combined with the information of other agents, we obtain global non-goal feature representation $e^{-g}_{global} = f_i(o^i_{-g}, \boldsymbol{o^{-i}}, \boldsymbol{a})$, where $\boldsymbol{o^{-i}}$ denotes observations except agent i and $\boldsymbol{a}$ is the joint action. The attention weight $w^{i,j}_{global}$ of goal j can be represented as:

$$w^{i,j}_{global} = \frac{\exp(e^{-g}_{global}e^j_{global})}{\sum_{j=1}^{L}\exp(e^{-g}_{global}e^j_{global})} \tag{4}$$

where $e^j_{global} = f_i(o^i_{g_j})$ is the feature representation of goal j. Then, the critic parameterized by ω_i of agent i evaluates actions more accurately based on goal cognition in the global view W^i_{global} :

$$Q^i(o, a; \omega_i) = f_i(e^i_{global}, x^i_{global}), \tag{5}$$

where $x^i_{global} = \sum_{j=1}^{L} w^{i,j}_{global}e^j_{global}$.

4.2 Training with Goal Consistency

Without a consistent goal, it is difficult to quickly emerge cooperative behavior from the learned goal cognition with limited knowledge. Therefore, we propose to train agents with two-level goal consistency, driving consistent goal cognition for effective exploration. First, we align the local view of goal cognition with the global view of goal cognition, which guides goal cognition learning in a self-supervised way. Second, we align the goal cognition of each agent with their teammates to foster a collective consistent stage-goal.

Consistency through Self-supervised Learning

Assuming that at time step t, the true goal $\mathcal{G}_t$ depends on the global state s_t represented as $\mathcal{G}_t = p(\mathcal{G}_t \mid s_t)$, where $p(\mathcal{G}_t \mid s_t)$ is the distribution of the true goal. To attain goal consistency, the goal $g^i_{t,local}$ selected by agent i in local view based on its local observation o^i_t should be similar to the true goal $\mathcal{G}_t$. Analogous to the consistency constraint using KL divergence in [Mao *et al.*, 2020], we achieve self-supervised learning by minimizing the following object:

$$\min KL(q(g^i_{t,local} \mid o^i_t) \| p(\mathcal{G}_t \mid s_t)) \tag{6}$$

Nevertheless, agents are not able to know the true distribution of goal in advance. To solve this problem, we suppose that the selected goal in a global view based on the joint observation $\mathbf{o}_t = \{o^1_t, \cdots, o^N_t\}$ and the joint action

$\mathbf{a}_t = \{a_t^1, \cdots, a_t^N\}$ of all agents will act as a precise approximation of the true stage-goal. Therefore, we replace the true goal $\mathcal{G}_t$ with the goal distribution $g^i_{t,global}$ attained in the critics to supervise the learning of goal cognition for the actors, and propose the self cognitive dissonance loss (SCD-loss) as:

$$\min KL(q(g^i_{t,local} \mid o^i_t) \| p(g^i_{t,global} \mid \mathbf{o}_t, \mathbf{a}_t)) \tag{7}$$

In such self-supervised way, the actors will continuously learn a better goal cognition.

Consistency through Aligning Teammates

Intuitively, if agents have similar goal cognition, they are more likely to reach the same goal. As mentioned before, each agent learns goal cognition in both the local view and global view, during which agents should keep consistent cognition with their teammates. We achieve this by aligning the goal cognition distribution with that of other agents.

Specifically, for agent i, we minimize the team cognitive dissonance loss (TCD-loss) in a local view:

$$\min \sum_{k \neq i} \mathrm{KL}(q(g^i_{t,\mathrm{local}} \mid o^i_t; \theta^i) \| p(g^k_{t,\mathrm{local}} \mid o^k_t; \theta^k)). \tag{8}$$

Similarly, the consistent global goal cognition is achieved by minimizing the team cognitive dissonance loss as

$$\min \sum_{k \neq i} \mathrm{KL}(q(g^i_{t,\mathrm{global}} \mid \mathbf{o}_t, \mathbf{a}_t; \omega^i) \| p(g^k_{t,\mathrm{global}} \mid \mathbf{o}_t, \mathbf{a}_t; \omega^k)) \tag{9}$$

Novelty of two-level goal consistency training. The two-level goal consistency training is rooted in goal cognition and encourages consistent exploration behaviors. Unlike the latest NCC-Q [Mao *et al.*, 2020] that learns aimless neighborhood cognition from low-level observations, goal consistency learning directly guides efficient exploration for agents at the goal level.

Training of MAGIC

MAGIC adopts an actor-critic architecture and can be easily combined with existing CTDE-based methods, such as MADDPG and Mean Field. Here, we illustrate the training process of MAGIC based on MADDPG, where the key is the two-level goal consistency training for driving cooperation. As shown in Fig.2, the two-level goal consistency training uses the team cognitive dissonance loss (TCD-loss) to constrain agents to be consistent with other agents in both local and global views of goal cognition, and exploits the self-cognitive dissonance loss (SCD-loss) to supervise the learning of goal cognition in local view.

Specifically, the critic $Q^i(o, a; \omega_i)$ is trained by minimizing the combination of the temporal-difference loss (TD-loss) and the team cognitive dissonance loss (TCD-loss) in global view as follows:

$$\mathcal{L}^{total}(\omega^i) = \mathcal{L}^{td}(\omega^i) + \alpha \mathcal{L}^{tcd}(\omega^i), \tag{10}$$

$$\mathcal{L}^{td}(\omega^i) = \mathbb{E}_{\mathbf{o},\mathbf{a},r_i,\hat{\mathbf{o}} \sim \mathcal{D}}[(Q^i(\mathbf{o}, \mathbf{a}; \omega^i) - y_i)^2], \tag{11}$$

$$\mathcal{L}^{tcd}(\omega^i) = \sum_{k \neq i} \mathrm{KL}(q(g^i_{global} \mid \mathbf{o}, \mathbf{a}; \omega^i) \| p(g^k_{global} \mid \mathbf{o}, \mathbf{a}; \omega^k)), \tag{12}$$

where $y_i = r_i + \gamma Q^i \left(\hat{\mathbf{o}}, \hat{\mathbf{a}}; \hat{\omega}^i\right)\big|_{\hat{a}^j = \pi(\hat{o}^j; \hat{\theta}^j)}$.

As for the actors of MAGIC, it is trained not only by the policy gradient but also the two-level goal consistency loss. First, we can write the gradient of the expected return for agent i as:

$$\nabla_{\theta^i} J(\theta^i) = \mathbb{E}_{\mathbf{o},\mathbf{a} \sim \mathcal{D}}[\nabla_{\theta^i} \pi^i(o^i; \theta^i) \nabla_{a^i} Q^i(\mathbf{o}, \mathbf{a}; \omega^i)|_{a_i = \pi^i(o^i; \theta^i)}]. \tag{13}$$

Then two-level goal consistency loss function of $\mathcal{J}^{tcd}(\theta^i)$ and $\mathcal{J}^{scd}(\theta^i)$ is

$$J^{tcd}(\theta^i) = \sum_{k \neq i} \mathrm{KL}(q(g^i_{local} \mid \mathbf{o}, \mathbf{a}; \theta^i) \| p(g^k_{local} \mid \mathbf{o}, \mathbf{a}; \theta^k)), \tag{14}$$

$$J^{scd}(\theta^i) = KL(q(g^i_{local} \mid o^i; \theta^i) \| p(g^i_{global} \mid \mathbf{o}, \mathbf{a}; \theta^i)). \tag{15}$$

Finally, the actor is updated by:

$$\nabla_{\theta^i} J^{total}(\theta^i) = \nabla_{\theta^i} J(\theta^i) + \lambda \nabla_{\theta^i} J^{tcd}(\theta^i) + \beta \nabla_{\theta^i} J^{scd}(\theta^i) \tag{16}$$

5 Experiments

In this section, we evaluate the effectiveness of MAGIC in three multistage cooperative tasks: resource collection, multipoint transport and cooperative endangered wildlife rescue.

5.1 Settings

Baselines

We compare MAGIC with independent DDPG [Lillicrap *et al.*, 2016], MADDPG [Lowe *et al.*, 2017], Mean Field [Yang *et al.*, 2018], MAAC [Iqbal and Sha, 2019], NCC-Q [Mao *et al.*, 2020] and GA-AC [Liu *et al.*, 2020]. MAAC focuses on information of the other agents dynamically by using the soft-attention (SA) mechanism. As we have found that DDPG shows better performance in our experiments than the original version using SAC, MAAC is implemented as DDPG+SA with experience sharing for a fair comparison. Mean Field MARL learns by approximating the interactions within agents with an average effect. NCC-Q learns neighborhood cognitive consistency to facilitate agent cooperation. GA-AC learns the relationship between agents by using game abstraction mechanism based on a two-stage attention network. In addition, we add sample efficient IGASIL [Hao *et al.*, 2019] as a baseline for the sparse reward task of endangered wildlife rescue.

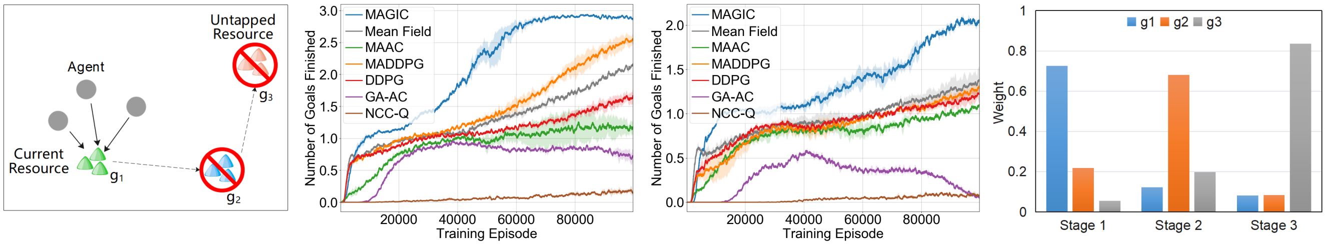
Training Details

For scenarios of resource collection and endangered wildlife rescue, we set $\alpha = 1$, $\lambda = 0.001$ and $\beta = 1$. For the multipoint transportation task, we set $\alpha = \lambda = 0.1$ and $\beta = 0.01$.

5.2 Resource Collection

Game Settings

The resource collection task requires agents to collect resources in multiple resource pools in sequence. There are N agents and L resource pools in a randomly generated environment, as shown in Fig.3(a). At different stage, one resource pool can be developed. The next target resource pool is accessible only when agents successfully mines the current target resource pool collaboratively. Therefore, agents should learn to recognize the goal of each stage and tap a resource pool cooperatively. Each agent receives a dense reward from the environment at each step, which are related to distance to the goals. We evaluate our method in the cases of $(N = 4, L = 3)$ and $(N = 10, L = 3)$.



(a) Resource collection (b) N=4,L=3 (c) N=10,L=3 (d) Goal cognition values

Figure 3: Resource collection: (a) Task description; (b)(c)The number of goals finished in the case of (N=4,L=3) and (N=10,L=3); (d) The average goal cognition values (attention weights) of agents at each stage when required to complete the goals $[g_1, g_2, g_3]$ in order.

Results

The results are shown in Fig.3(b) and Fig.3(c), where we evaluate the number of goals finished. Obviously, learning with shaped rewards reduces the difficulty of exploration for all methods, MAGIC outperforms other baselines in accelerating the learning process and improving the policy performance. Moreover, in Fig.3(c), as the input dimension increases exponentially with the number of agents, MAGIC still learns faster than the other methods with higher performance while the performance of MADDPG decreases sharply, indicating that goal consistency promotes the efficient exploration and cooperation of large-scale agents. The other baselines that focus on neighborhood-level coordination perform poorly in global objective optimization, even worse than MADDPG that learns global-level coordination by naively considers information of all agents. NCC-Q is found ineffective to solve complex multistage tasks, since agents hardly benefit from the aimless neighborhood cognitive consistency, which even introduces complex network training process. The poor performance of GA-AC and MAAC also implies that making much effort on learning relationship among agents hinders efficient exploration of goals in multistage tasks.

To further analyze whether agents learn to cognitive goals of different stages and maintain consistent goals, we show the average goal cognition values (attention weights) of all agents in Fig.3(d), which describes the changes of average attention weights over three stages in the task that requires to complete goals in the sequence of $[g_1, g_2, g_3]$. We observe that the goal with the highest average weight at each stage corresponds to the actual goal, indicating that agents have good goal cognition. Moreover, the weights of the selected goals are clearly distinguished from that of other goals, which means that the agents have formed a consistent goal cognition.

5.3 Multi-Point Transportation

Game Settings

The multi-point transportation task requires N agents to cooperatively push a big ball to L destinations in order, as shown in the Fig.4(left). This task is similar to the real-world cargo transportation task that require the cargo to pass through multiple delivery points in sequence according to the prescribed routes. Dense rewards are offered to agents in each step. We evaluate our method in the case of $(N = 4, L = 3)$.

Results

The multi-point transportation task is more difficult than the resource collection task, due to the requirement for pushing a ball cooperatively and navigating to multiple points. In Fig.4(right), we show the learning curves of the number of goals finished by different algorithms. MAGIC learns faster and converges to a much higher goal completion rate than all the baselines. Without a consistent goal, agents trained by other methods may push the ball in different directions, thus wasting a lot of exploration time. However, agents trained by MAGIC learn to quickly transfer the moving direction to the next goal when they finish the current stage-goal. Moreover, when reaching the last destination, some agents will turn to the opposite direction of the ball to slow it down. It is verified that agents have benefited from goal consistency training, which greatly accelerates cooperation among agents and promotes effective exploration.

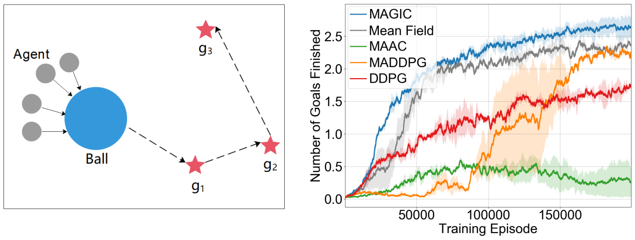


Figure 4: Multi-point transportation: (*Left*) Task description. (*Right*) The number of goals finished for MAGIC and the baselines.

5.4 Endangered Wildlife Rescue

Game Settings

The endangered wildlife rescue is a sparse reward task modified from the predator-prey scenario in [Lowe *et al.*, 2017]. As shown in the Fig.5(left), there are $N = 2$ slowly rescue agents and $M = 2$ faster wounded animals. One of the animal g_1 is seriously injured (in red) which should be rescued first, and the other animal g_2 in pink is less injured. The target for each rescue agent is to rescue all wounded animals according to the severity of the injury. Agents are rewarded only if they save an animal simultaneously, and punished for miss-coordination. Different wounded animals correspond to different rewards and different risks (penalties). The sparse reward setting poses significant challenges to exploration and policy learning.

Results

To make different algorithms comparable, we pre-train both the rescue agents and wounded animal agents with DDPG and save the animal models during training. Then, we reuse the same pre-trained animal models as the default policies for the

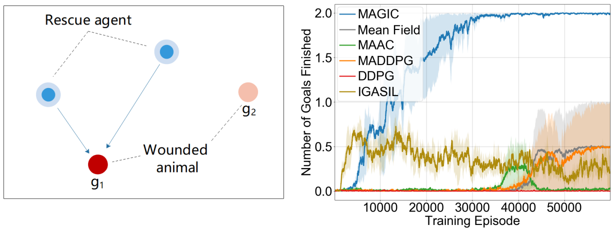


Figure 5: Endangered wildlife rescue: (*Left*) Task description. (*Right*) The number of goals finished for MAGIC and other five baselines.

wounded animals in all experiments, thus the environment is fully cooperative.

As can be seen in Fig.5(right), MAGIC significantly surpasses all baselines in convergence speed and final performance. In this difficult sparse reward task, MADDPG and Mean Field perform poorly in learning efficiency and result in sub-optimal policies. IGASIL also fails to achieve continually cooperative by aimlessly self-imitation learning. While these baseline methods hardly learn to rescue animals cooperatively by aimless policy learning, the advantage of MAGIC is more prominent in the multistage task with sparse reward setting, where consensus is extremely critical due to the lack of effective guiding signals.

5.5 Discussion

Ablation Study

In this section, we further verify how goal consistency affects the learning process in the endangered wildlife rescue task. We implement three ablation variants of MAGIC: (1) MAGIC w/o TCD-loss removes TCD-loss during training, (2) MAGIC w/o SCD-loss removes SCD-loss, (3) GON removes both TCD-loss and SCD-loss.

As shown in Fig.6(left), removing TCD-loss (MAGIC w/o TCD-loss) results in lower learning efficiency than removing SCD-loss (MAGIC w/o SCD-loss), which means that aligning with teammates plays a more important role in driving goal consistency than self-supervised learning. MAGIC significantly outperforms GON in final goal completion rate by exploiting the two-level goal consistency training. Although GON fails to coordinate to rescue both wounded animals, it performs better than all the baseline methods shown in the Fig.5(right), indicating that learning goal cognition before making a decision provides effective guidance for agent training. More importantly, the superiority of MAGIC demonstrates that the two-level goal consistency loss further accelerates goal-oriented exploration process of agents.

Universality Study

To study the universality of MAGIC, we increase the coordination difficulty and the number of agents in the resource collection task to investigate the scope of solvable problems by using the MAGIC framework.

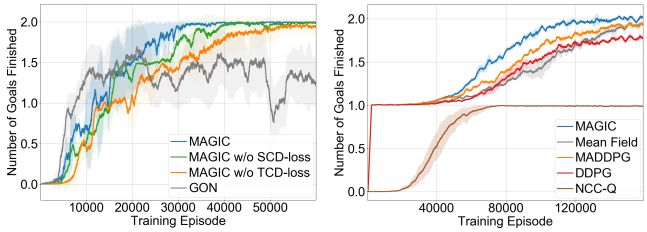
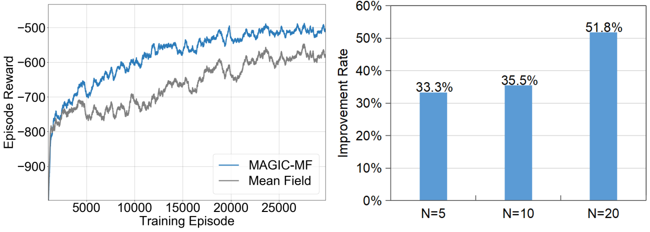


Figure 6: (*Left*) Ablation results in the endangered wildlife rescue task. (*Right*) Results in a two-stage resource collection task with two resource pools in the second phase.



(a) (b)

Figure 7: Results of resource collection tasks ($L = 3$) with large-scale agents. (a) The episode rewards for MAGIC-MF and Mean Filed algorithm when $N = 20$; (b) Percentage of sample efficiency that MAGIC-MF improves compared to Mean Field algorithm when $N = [5, 10, 20]$. Note that parameter sharing trick is used to ease training in the universality study experiments.

On the one hand, considering the coordination requirements of the actual task, we build a two-stage resource collection task with two resource pools in the second phase. Therefore, the optimal strategy of agents in the second stage is to divide into two groups coordinately. As shown in Fig.6(right), MAGIC still achieves higher sample efficiency compared with other algorithms, which implies that MAGIC can effectively accelerate the learning process of cooperative tasks, even in tasks with coordination difficulty. We also observe that all algorithms converge to suboptimal strategies, where MAGIC agents tend to act concertedly under the constraints of goal consistency. To study how to divide groups and impose goal consistency constraints among agents in the same group is considerable future work.

On the other hand, considering complex scenarios of large-scale agents, we combine MAGIC with Mean Field algorithm to form MAGIC-MF and evaluate it in the case of $N = [5, 10, 20]$. Although the Mean Field algorithm has been proved to be capable of handling large-scale agents in single-stage tasks, its aimless exploration in multi-stage tasks leads to difficulty in finding stage goals. However, MAGIC-MF significantly improves sample efficiency, especially in large-scale agents scenarios (as shown in Fig.7). It suggests that MAGIC can be easily combined with existing advanced algorithms to accelerate learning in a variety of complex multi-stage cooperative tasks.

6 Conclusions

In this paper, we focus on multistage cooperative tasks with sequential goals and propose multi-agent goal consistency (MAGIC) framework, which improves goal cognition ability for agents and achieves effective exploration. MAGIC learns goal cognition by using a goal-oriented actor-critic model, which enables agents to focus on different goals in different stages adaptively . Moreover, MAGIC further facilitates exploration efficiency by introducing tow-level goal consistency. Empirically results demonstrate that MAGIC improves sample efficiency in challenging multistage tasks universally, especially in tasks with sparse reward and large-scale agents.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (Grant No.2020YFB2104000), the National Natural Science Foundation of China (Grant Nos. 62172154, 61772559), the Hunan Provincial Natural Science Foundation of China under grant No.2020JJ3016. Dr. Xuan Liu is partially supported by the National Defense Science and Technology Innovation Special Zone Project of China. Xuan Liu is the corresponding author of this paper.

References

[Ding *et al.*, 2020] Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning individually inferred communication for multi-agent cooperation. In *Proceedings of Conference on Neural Information Processing Systems*, 2020.

[Gronauer and Diepold, 2021] S. Gronauer and K. Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, pages 1–49, 2021.

[Hao *et al.*, 2019] Xiaotian Hao, Weixun Wang, Jianye Hao, and Yaodong Yang. Independent generative adversarial self-imitation learning in cooperative multiagent systems. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pages 1315–1323, 2019.

[Iqbal and Sha, 2019] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2961–2970, 2019.

[Jiang and Lu, 2018] Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. In *Proceedings of Conference on Neural Information Processing Systems*, pages 7265–7275, 2018.

[Li *et al.*, 2021a] Chenghao Li, Chengjie Wu, Tonghan Wang, Jun Yang, Qianchuan Zhao, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. *CoRR*, abs/2106.02195, 2021.

[Li *et al.*, 2021b] Wenhao Li, Xiangfeng Wang, Bo Jin, Junjie Sheng, Yun Hua, and Hongyuan Zha. Structured diversification emergence via reinforced organization control and hierarchical consensus learning. *CoRR*, abs/2102.04775, 2021.

[Lillicrap *et al.*, 2016] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

[Lin *et al.*, 2021] Toru Lin, Minyoung Huh, Chris Stauffer, Ser-Nam Lim, and Phillip Isola. Learning to ground multi-agent communication with autoencoders. *CoRR*, abs/2110.15349, 2021.

[Liu *et al.*, 2020] Yong Liu, Weixun Wang, Yujing Hu, Jianye Hao, Xingguo Chen, and Yang Gao. Multi-agent game abstraction via graph attention neural network. In *Proceeding of Conference on Artificial Intelligence*, pages 7211–7218, 2020.

[Liu *et al.*, 2021] Iou-Jen Liu, Unnat Jain, Raymond A. Yeh, and Alexander G. Schwing. Cooperative exploration for multi-agent deep reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6826–6836, 2021.

[Lowe *et al.*, 2017] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceeding of Neural Information Processing Systems*, pages 6379–6390, 2017.

[Mahajan *et al.*, 2019] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. MAVEN: multi-agent variational exploration. In *Proceedings of Conference on Neural Information Processing Systems*, pages 7611–7622, 2019.

[Mao *et al.*, 2019] Hangyu Mao, Zhengchao Zhang, Zhen Xiao, and Zhibo Gong. Modelling the dynamic joint policy of teammates with attention multi-agent DDPG. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1108–1116, 2019.

[Mao *et al.*, 2020] Hangyu Mao, Wulong Liu, Jianye Hao, Jun Luo, Dong Li, Zhengchao Zhang, Jun Wang, and Zhen Xiao. Neighborhood cognition consistent multi-agent reinforcement learning. In *Proceeding of Conference on Artificial Intelligence*, pages 7219–7226. AAAI Press, 2020.

[Ryu *et al.*, 2020] Heechang Ryu, Hayong Shin, and Jinkyoo Park. Multi-agent actor-critic with hierarchical graph attention network. In *Proceeding of Conference on Artificial Intelligence*, pages 7236–7243, 2020.

[Sukhbaatar *et al.*, 2016] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Proceedings of Conference on Neural Information Processing Systems*, pages 2244–2252, 2016.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2048–2057, 2015.

[Yang *et al.*, 2018] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5567–5576, 2018.

[Yang *et al.*, 2020] Jiachen Yang, Alireza Nakhaei, David Isele, Kikuo Fujimura, and Hongyuan Zha. CM3: cooperative multi-goal multi-stage multi-agent reinforcement learning. In *Proceeding of the 8th International Conference on Learning Representations*, 2020.