PKU-IAI Technical Report: TR-PKU-IAI-2024-0054

Planning and Modeling: A Fast Few-Shot Adaptation Method of Planning with Game Theory of Mind

Liangyu Wu
Yuanpei College
Peking University
wuliangyu@stu.pku.edu.cn

Yizhe Huang
Institute for Artificial Intelligence
Peking University
szhyz@pku.edu.cn

Xue Feng
Institute for Artificial Intelligence
Peking University
fengxue@bigai.ai

Abstract

Recent research in artificial intelligence has been increasingly focused on the development of autonomous agents that can effectively interact with and adapt to other agents. Multi-agent reinforcement learning (MARL) is a prominent set of algorithms that address the adaptation challenge. Although there has been significant progress in scenarios requiring many interactions, the field still struggles with efficient adaptation in few-shot learning situations. In this paper, we propose a novel method for rapid adaptation in mixed-motive games, which we call Planning with Theory of Mind (PToM). This approach incorporates modeling other agents within game contexts and planning actions based on information and beliefs derived from Theory of Mind (ToM). PToM consists of two integrated modules: an opponent modeling module that infers other agents' objectives and learns their policies using ToM, and a planning module that applies a modified Monte Carlo Tree Search (MCTS) to find the most effective response. Our method enhances adaptation efficiency by updating beliefs about other agents' goals both during and between episodes, and by utilizing insights from the opponent modeling to inform the planning process. The introduction of PToM contributes to a greater understanding of the development of social intelligence in complex multi-agent environments.

1 Introduction

Constructing agents capable of rapidly adapting to previously unseen agents remains a longstanding challenge in artificial intelligence. This ability, known as few-shot adaptation, has been studied extensively in zero-sum games and common-interest environments using well-performed Multi-Agent Reinforcement Learning (MARL) algorithms. However, most realistic multi-agent decision-making scenarios extend beyond these predefined competitive or cooperative relationships. Instead, they should be abstracted as mixed-motive environments, where agent interactions are non-deterministic, and an agent's optimal responses may change based on others' behavior. Failing to adapt quickly to

new opponents in such environments can harm not only the focal agent's interests but also the entire group's welfare.

The Challenge in Mixed-Motive Environments While MARL algorithms have succeeded in zero-sum and pure-cooperative environments, they struggle in mixed-motive settings due to their reliance on reward-specific techniques like minimax, Double Oracle, or IGM condition. These techniques are not directly applicable in mixed-motive scenarios, where non-deterministic relationships between agents and general-sum reward structures pose additional challenges for decision-making and few-shot adaptation.

Inspiration from Cognitive Psychology Drawing inspiration from cognitive psychology, which emphasizes hierarchical cognitive mechanisms, we propose an algorithm called Planning with Theory of Mind (PToM). This hierarchical approach unifies high-level goal reasoning with low-level action planning. In machine learning research, hierarchical goal-directed planning has been proven effective for few-shot problem-solving. PToM comprises two modules: an opponent modeling module and a planning module. The opponent modeling module infers opponents' goals and learns their goal-conditioned policies using Theory of Mind (ToM), which involves understanding others' mental states (goals and beliefs) from their actions. To enhance inference efficiency, beliefs about others' goals are updated both between and within episodes. The information from the opponent modeling module is then used by the planning module, based on Monte Carlo Tree Search (MCTS), to compute the next action.

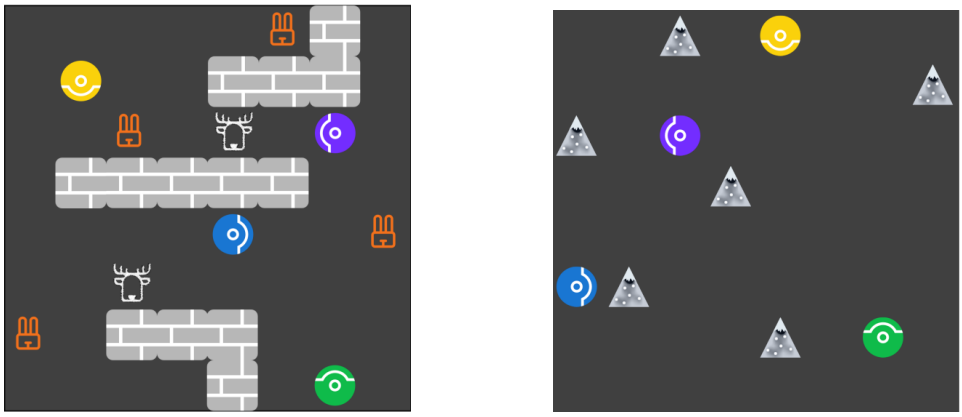
Experimental Evaluation We evaluate PToM's few-shot adaptation ability in the Markov Stag-Hunt (MSH) Game, extending classic paradigms in game theory. MSH demonstrates how an agent's best response in a mixed-motive environment depends on opponents' strategies. Our experimental results reveal that PToM outperforms baselines, including established MARL algorithms (LOLA, social influence, A3C, prosocial-A3C, PR2, and direct-OM). Furthermore, PToM achieves high rewards in self-play, showcasing exceptional decision-making abilities in mixed-motive games. Notably, we observe the emergence of social intelligence through interactions among multiple PToM agents, including self-organized cooperation and alliances among disadvantaged agents.

2 Related work

In the realm of multi-agent reinforcement learning (MARL), researchers have delved into the complexities of mixed-motive games. One prevalent approach involves augmenting intrinsic rewards to encourage collaboration and consideration of the impact on other agents, alongside the pursuit of extrinsic rewards. Noteworthy examples include ToMAGA (Nguyen et al., 2020), MARL with inequity aversion (Hughes et al., 2018), and prosocial MARL (Peysakhovich et al., 2018). However, many of these algorithms rely on hand-crafted intrinsic rewards and assume access to other agents' rewards. Unfortunately, this reliance can render them exploitable by self-interested algorithms and less effective in realistic scenarios where others' rewards remain hidden (Komorita et al., 1995; Yoshida et al. [4], 2008)

To address these challenges, Jaques et al. (2019) introduced intrinsic social influence rewards, leveraging counterfactual reasoning to assess an agent's actions' impact on opponents' behavior. Another notable model is LOLA (Foerster et al., 2018), along with its extensions like POLA (Zhao et al., 2022) and M-FOS (Lu et al., 2022). Unlike traditional approaches that treat other agents as static components of the environment, LOLA takes the impact of an agent's learning process into consideration. However, LOLA requires knowledge of opponents' network parameters, which may not be feasible in many real-world scenarios. To mitigate this problem, LOLA with opponent modeling relaxes the parameter requirement, but scalability issues may arise in complex sequential environments necessitating long action sequences for reward accumulation (Leibo et al. [3], 2021)

Albrecht and Stone [1] have provided a comprehensive review of an extensively-explored field opponent modeling, with which our work aligns. A typical framework in this domain is I-POMDP (Gmytrasiewicz & Doshi, 2005), which maintains dynamic beliefs over the physical environment and beliefs about other agents' intentions. I-POMDP maximizes a value function based on these beliefs to determine the next action. However, the nested belief inference in I-POMDP suffers from severe computational complexity, limiting its practicality in complex environments. In contrast, our proposed approach, Planning with Theory of Mind (PToM), explicitly leverages beliefs about other agents' goals and policies. PToM employs a neural network model of other agents (MOA) and an



(a) **MSH**. (b) **MSG**.

Figure 1: **Environment Visualization for MSH and MSG games.**

efficient Monte Carlo Tree Search (MCTS) planner for sequential decision-making, avoiding the pitfalls of nested belief inference.

Theory of mind (ToM), originally rooted in cognitive science and psychology (Baron-Cohen et al., 1985), has undergone transformation into computational models over the past decade. These models infer agents' mental states, including goals and desires. While Bayesian inference has been a popular technique for making ToM computational (Baker et al., 2011; Poppel & Kopp, 2018; Wu et al., 2021; Zhi-Xuan et al., 2022), recent advances in neural networks have also explored achieving ToM (Rabinowitz et al., 2018; Shu & Tian, 2018; Wen et al., 2019; Moreno et al., 2021). PToM provides a practical and effective framework for utilizing ToM, extending its application to mixed-motive environments where both competition and cooperation play pivotal roles, and opponents' goals remain private and volatile.Yoshida et al. [4]

Lastly, Monte Carlo Tree Search (MCTS) stands as a widely adopted planning method for optimal decision-making. While recent breakthroughs like AlphaZero (Silver et al., 2018) and MuZero (Schrittwieser et al., 2020) have harnessed MCTS as a general policy improvement operator alongside neural network-based policies, its scalability remains a concern in multi-agent environments. The joint action space grows exponentially with the number of agents, but our approach mitigates this challenge by estimating opponent policies and focusing planning efforts solely on the focal agent's actions.Huang et al. [2]

3 Problem formulation

In our research, we delve into multi-agent decision-making within mixed-motive environments, which we model as Markov games (Littman, 1994). Let's break down the key components of our formulation:

Agent Setup We consider a set of agents, denoted by $N = \{1, 2, \ldots, n\}$. Each agent, indexed by i, selects actions from its action space $A_i = \{a_i\}$.

The joint action space A is defined as $A_1 \times A_2 \times \ldots \times A_n$, where a joint action $a_{1:n}$ leads to a state transition based on the transition function $T : S \times A \times S \to [0, 1)$. Specifically, after agents collectively take action $a_{1:n}$, the environment transitions from state s to s' with probability $T(s'|s, a_{1:n})$.

The immediate reward received by agent i after taking joint action $a_{1:n}$ in state s is denoted by the reward function $R_i : S \times A \to \mathbb{R}$.

We introduce a discount factor γ to account for future rewards, and T_{max} represents the maximum episode length.

Agent i's policy, denoted as $\pi_i : S \times A_i \to [0, 1]$, specifies the probability $\pi_i(a_i|s)$ of choosing action a_i in state s.

Goals and Beliefs Our study environments involve a set of goals, represented by $G = G_1 \times G_2 \times \cdots \times G_n$. Each G_i corresponds to the set of goals for agent i.

Interestingly, an agent cannot directly access another agent's true goal. However, agent i can infer agent j's goal based on its observed action sequence.

To capture this inference process, agent i maintains a belief over agent j's goals, $b_{ij} : G_j \to [0, 1]$. This belief serves as a probability distribution over possible goals for agent j.

Evaluation Metrics We evaluate algorithms in terms of self-play and few-shot adaptation within mixed-motive environments. Multiple agents use the same algorithm and train from scratch. The performance is assessed by the expected reward after convergence, demonstrating the algorithm's autonomous decision-making abilities.Algorithms must recognize and respond effectively to unknown policies within a limited number of episodes. Their performance is measured by the rewards achieved during these brief interactions.

Our research aims to shed light on how agents can observe and adapt to their counterparts in mixed-strategy games, such as the stag hunt game. By understanding the interplay between individual interests and collaborative dynamics, we can uncover valuable insights for designing robust multi-agent systems.

4 Methods

In this section, we introduce Planning with Theory of Mind (PToM), a novel algorithm designed for multi-agent decision-making within mixed-motive environments. PToM comprises two essential modules: an opponent modeling module responsible for inferring opponents' goals and predicting their behavior, and a planning module that guides the focal agent in determining the best response based on information inferred from the opponent modeling module.

In the Opponent Modeling Module, drawing inspiration from cognitive psychology, which posits that others' behavior is goal-directed (Gergely et al., 1995; Buresh & Woodward, 2007), and that agents exhibit stability in pursuit of specific goals (Warren, 2006), our opponent modeling module operates with a two-level hierarchy. We employ ToM to infer opponents' internal goals by analyzing their action sequences. Based on the inferred goals and the current environment state, the low-level component learns goal-conditioned policies to model opponents' atomic actions.

In the planning module, we utilize Monte Carlo Tree Search (MCTS) to determine the focal agent's optimal response, considering the uncertainty surrounding opponents' goals. We handle this uncertainty by sampling Opponent Goal Combination. Given the estimated opponent policies from the opponent modeling module, we sample multiple combinations of opponent goals according to the current belief. We then compute the action that maximizes the average return over these sampled configurations.

Following the principles of AlphaZero (Silver et al., 2018) and MuZero (Schrittwieser et al., 2020), we maintain a policy and a value network to enhance MCTS planning. The planned action and its value update the neural network.

4.1 Goal Inference Challenges

While PToM effectively summarizes opponents' objectives based on interaction history (as depicted in Figure 2), it faces challenges related to potential changes in opponents' goals within episodes. To address this, we propose two update procedures based on ToM:

Intra-ToM (Within-Episode Inference) Intra-ToM infers an opponent's immediate goals within a single episode (K) based on their past trajectory. This ensures that PToM can quickly adapt to in-episode behavior changes by updating agent i's belief about agent j's goals at time t, denoted as $b_{i,j}^{K,t}(g_j)$.

Inter-ToM (Across-Episode Inference) Inter-ToM summarizes opponent goals based on historical episodes. The belief update between adjacent episodes is defined by a time-discounted modification of the Monte Carlo estimate, considering the importance of historical information (controlled by the horizon weight α).

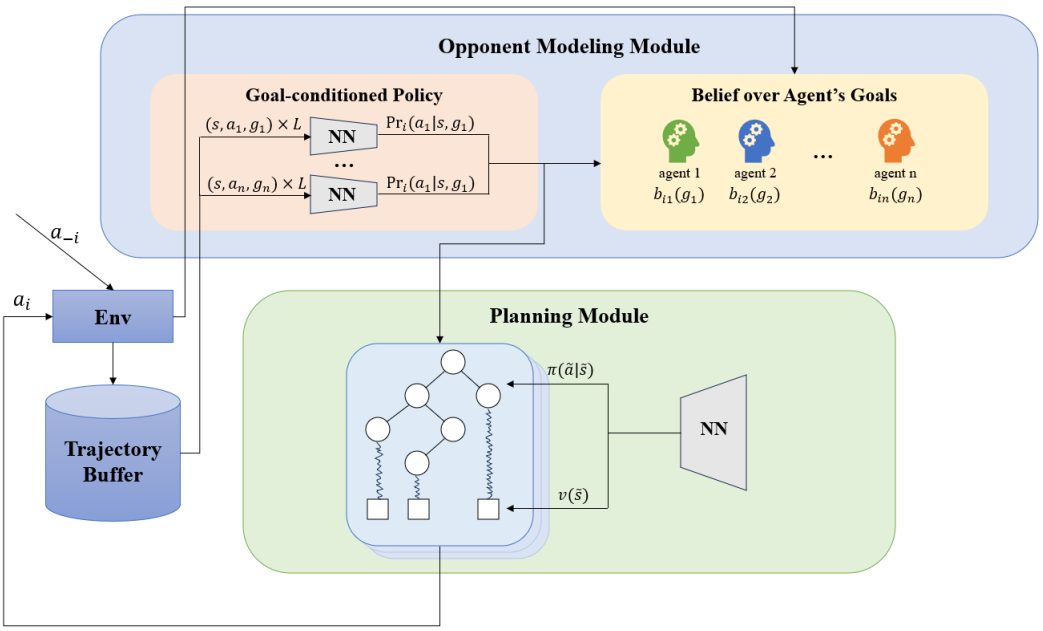


Figure 2: Overview structure of PToM

4.2 Neural Network-Based Goal-Conditioned Policies

To obtain the goal-conditioned policy, we use a neural network ω. Training involves collecting sets of $(s^j_{K,t}, a^j_{K,t}, g^j_{K,t})$ from episodes and updating ω to minimize the negative log-likelihood.

Our approach leverages MCTS to compute advantageous actions, even when opponent policies contain uncertainty over their goals. By sampling opponent goal combinations and balancing computational complexity, we ensure effective planning.

5 Experiments

5.1 Baseline Algorithms and Few-Shot Adaptation

In the context of evaluating the performance of the PToM (Planning To Model) algorithm, we introduce several baseline algorithms. These baselines serve a dual purpose during the evaluation of few-shot adaptation. Firstly, they act as unfamiliar opponents to test PToM's ability to adapt in a few-shot scenario. Secondly, we evaluate the few-shot adaptation ability of these baseline algorithms to demonstrate PToM's superiority.

The baseline algorithms include the following:

- LOLA: These agents consider a 1-step look-ahead update of opponents and update their own policies based on the updated policies of opponents.
- SI (Social Influence): These agents have an intrinsic reward term that incentivizes actions maximizing their influence on opponents' actions, accessed through counterfactual reasoning.
- A3C (Asynchronous Advantage Actor-Critic): These agents are trained using a well-established reinforcement learning (RL) technique. A3C agents update their policies asynchronously.
- PS-A3C (Prosocial-A3C): These agents share rewards between players during training, optimizing the per-capita reward instead of individual rewards, emphasizing cooperation.
- PR2: These agents model how opponents would react to their potential behaviors, allowing them to find the best response.

Additionally, we construct some rule-based strategies, including random policies, cooperators (agents consistently adopting cooperative behavior), and defectors (agents consistently adopting exploitative

behavior). In specific environments (e.g., MSH and MSG), the goals of cooperators and defectors vary (e.g., hunting stags or hares).

The experiment consists of two phases. i) Self-Play: Agents are trained until convergence, and their self-play performance (measured by average reward) demonstrates their ability to achieve cooperation. ii) Few-Shot Adaptation: A focal agent interacts with three opponents using different algorithms. The focal agent's average reward during the final 600 steps measures its few-shot adaptation ability. Parameters at the start of the adaptation phase are derived from self-play.

Notably, PToM shows effective adaptation, cooperating with other agents and adjusting to non-cooperative scenarios when facing opponents with fluctuating strategies.

5.2 Adaptation in Specific Environments

- MSH-4h1s: In this environment, only PToM, direct-OM, and PS-A3C learn the strategy of hunting stags. PS-A3C, however, may not effectively learn the relationship between hunting and rewards, resulting in inferior overall rewards. LOLA, SI, and A3C exhibit different strategies, while PR2 fails to work due to the environment's dynamics.

- MSH-4h2s: All algorithms learn the strategy of cooperatively hunting stags. PToM and A3C yield higher returns, while PS-A3C tends to delay hunting. PToM's performance surpasses other algorithms when facing dynamically adjusting opponents.

5.3 PToM's Efficient Adaptation

PToM's ability to adapt efficiently to unseen agents is exemplified in scenarios like facing three defectors (always attempting to hunt the nearest hare). By discerning real-time opponent goals, PToM adjusts its policy accordingly, achieving substantial strategic adaptations.

6 Conclusion and discussion

Over the past three months, I have been extensively participating in the PToM project. As a green hand, I actively learned RL, MAS and Game Theory as preliminary knowledge to this project. I watched many online lessons and attended professional lectures and lessons, which broadened my horizons a lot. And I am deeply exposed to the environment of scientific research, which will possibly benefit me in my whole life. I would like to express my sincere gratitude to my supervisor and my supervising doctoral student.

Our project has been focusing on Planning with Theory of Mind (PToM), an innovative hierarchical algorithm designed for few-shot adaptation to previously unseen opponents in mixed-motive environments. The core components of PToM include an opponent modeling module, which infers opponents' goals and behavior, and a planning module that generates the focal agent's best response based on the inferred information. Empirical results demonstrate that PToM outperforms state-of-the-art multi-agent reinforcement learning (MARL) algorithms, particularly in handling mixed-motive scenarios during self-play and adapting to novel opponents.

While PToM exhibits remarkable capabilities, several limitations guide our future work. The first one is about the goal definition in diverse environments. PToM requires clear goal definitions in any environment. To enhance its ability to generalize across various scenarios, we need techniques that autonomously abstract goal sets. This abstraction process should adapt to different contexts and scenarios seamlessly. Secondly, human behavior modeling needs to be improved. Our investigations primarily involve well-established algorithms as opponents. However, none of these models adequately capture human behavior. It would be intriguing to explore how PToM performs in few-shot adaptation scenarios involving human participants. Given that PToM is inherently self-interested, there's a risk that it may not always make decisions aligned with human interests. One potential solution is to leverage PToM's ability to infer and optimize for human values and preferences during interactions, thereby assisting humans in complex decision-making environments.

References

[1] Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018. 2

[2] Yizhe Huang, Anji Liu, Fanqi Kong, Yaodong Yang, Song-Chun Zhu, and Xue Feng. Planning with theory of mind for few-shot adaptation in sequential social dilemmas. 2023. 3

[3] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. 139:6187–6199, 18–24 Jul 2021. 2

[4] Wako Yoshida, Ray J Dolan, and Karl J Friston. Game theory of mind. *PLoS computational biology*, 4(12):e1000254, 2008. 2, 3