Towards Spontaneous Cooperation in Multi-Agent Reinforcement Learning using Explicit Goal Recognition

Anonymous authors
Paper under double-blind review

Abstract

Spontaneous cooperation —– the ability to assist others without explicit instruction or coordination —– is a hallmark of intelligent social behavior observed in humans and other animals. However, most Multi-Agent Reinforcement Learning (MARL) approaches lack mechanisms for intuitive, goal-directed helping due to limited modeling of other agents' internal states. In this paper, we explore a Theory of Mind (ToM)-inspired approach to address this gap, enabling artificial agents to infer and support the hidden goals of their teammates. Building on the Hidden Goal Markov Decision Process (HGMDP) framework, we introduce a baseline evaluation in a simplified collaborative domain in which an assistant agent must infer whether a leader agent is hungry or thirsty and deliver the appropriate item without direct communication. This preliminary system demonstrates how basic goal inference can enable spontaneous, context-sensitive cooperation. These findings lay the groundwork for future development of MARL agents capable of adaptive, intuitive assistance in more complex environments.

1 Introduction

Despite the obvious interdependencies that exist between agents, existing work in Multi-Agent Reinforcement Learning (MARL) typically assumes that each agent operates with limited or no explicit representation of other agents' internal states, goals, or learning processes (Albrecht and Stone, 2018). In contrast, multi-agent systems grounded in the beliefs-desires-intentions (BDI) paradigm emphasize the importance of modeling intentions as a means to support coherent and effective collaboration (Rao *et al.*, 1995; Grosz and Kraus, 1996). The disconnect between these two perspectives leaves a gap in our ability to develop MARL agents that can engage in fluid, human-like teamwork.

To bridge this gap, we propose a theory of mind (ToM)-inspired approach that enables reinforcement learning agents to explicitly infer the goals or mental states of their teammates to improve coordination and collaborative performance (Georgeff *et al.*, 1999; Langley *et al.*, 2022). Our focus is on reducing the cognitive and computational overhead needed for an agent to act helpfully, particularly in environments where spontaneous cooperation is essential. This line of work is grounded in observations from cognitive science, where even toddlers and chimpanzees can display forms of intuitive, goal-directed helping behavior—such as assisting someone in opening a cabinet—without prior training or explicit communication (Warneken and Tomasello, 2006).

To begin tackling the challenge of such spontaneous cooperation in artificial agents, we build upon the Hidden Goal Markov Decision Process (HG-MDP) framework (Fern *et al.*, 2014), which provides a natural formalism for modeling goal ambiguity in interactive settings. Specifically, we introduce a simple yet illustrative collaborative environment inspired by the popular Overcooked domain—a benchmark used extensively in MARL research due to its structured tasks, flexible agent roles, and rich coordination challenges. Within this domain, we implement a hungry-thirsty setting, where a leader agent may be either hungry or thirsty, and an assistant agent must infer and support the leader's latent goal by delivering the correct item (e.g., sushi or water). This setting is particularly challenging for non-adaptive agents, as effective assistance requires dynamic goal recognition and context-sensitive action selection.

We present a baseline system in which the assistant agent attempts to infer the leader's hidden goal and act supportively, without relying on explicit communication or complex mental simulation. Rather than introducing a new recognition or cooperation algorithm, this work offers a simple baseline and a controlled, clean environment to evaluate goal recognition and cooperative behavior both independently and in combination. Through this setup, we evaluate how simple forms of goal inference can enhance cooperative behavior, even in cases where the assisting agent lacks access to the leader's internal policy or reward function. This paper reports on our preliminary findings, highlights the potential of cognitively inspired goal recognition in MARL, and outlines key avenues for enhancing the reasoning and interaction capabilities of assisting agents in future work.

2 Background

2.1 Multi-Agent Reinforcement Learning (MARL)

A single-agent sequential decision process is modeled as a Markov Decision Process (MDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$. At each time step t, the agent observes a state $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}$ according to its policy $\pi(a|s)$. The environment provides a reward r_t and transitions to a new state s_{t+1} based on the transition function $T(s_{t+1}|s_t, a_t)$. The agent aims to learn a policy that maximizes the expected return $G = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, where $\gamma \in [0, 1)$ is the discount factor.

In Multi-Agent Reinforcement Learning (MARL), the environment is typically modeled as a multi-agent MDP (also known as a Markov game) for n agents, represented by the tuple $\langle S, A_1, \ldots, A_n, T, R, \gamma \rangle$. Here, S denotes the set of joint environment states; $A_1, \ldots, A_n$ represent the action sets available to each agent; T is the state transition function based on the joint state and the agents' actions, $T(s_{t+1}|s_t, a_{1,t}, \ldots, a_{n,t})$; and γ is the discount factor. In a fully cooperative setting, $R : S \times A_1 \times \cdots \times A_n \to \mathbb{R}$ is the team's shared reward function. This formulation also assumes that agents have full observability and thus each has the same state. However, the full transition and reward functions may not be available to the agents. We consider two classes of MARL algorithms or settings: those without Theory of Mind (ToM) and those incorporating ToM.

MARL Without Theory of Mind In settings without ToM, agents typically do not model the internal states or intentions of other agents. A common approach is independent Q-learning (Tan, 1993), where each agent treats other agents as part of the environment and selects actions based solely on its own observations. Learned information is not shared between agents. Centralized Training with Decentralized Execution (CTDE) frameworks, such as QMIX (Rashid *et al.*, 2020) and MADDPG (Lowe *et al.*, 2017), allow agents to act independently during execution but share information during training. This setup enables agents to pool knowledge and learn more efficiently, often resulting in identical policies across agents. Model-based RL approaches, like R-max (Brafman and Tennenholtz, 2003), can be advantageous when the transition and reward functions are known or can be learned. Agents can compute joint policies through planning methods like dynamic programming, even without explicit communication. For discrete state and action spaces, tabular learning methods may suffice, while continuous or complex tasks often necessitate function approximation techniques.

MARL With Theory of Mind Incorporating ToM into MARL involves agents modeling the beliefs, intentions, or learning processes of other agents. Ad-hoc teamwork is a related problem where one can control only a single agent, while teammates may have different capabilities and learning abilities (Mirsky *et al.*, 2022). For example, Ribeiro *et al.* (2022) present a Bayesian online prediction algorithm for ad hoc teamwork under partial observability (ATPO), enabling agents to collaborate with unknown teammates on unknown tasks without prior coordination, observable teammate actions, or environmental rewards. Learning with Opponent-Learning Awareness (LOLA) is a method where an agent anticipates and influences the learning updates of other agents by explicitly modeling their learning processes (Foerster *et al.*, 2018; Zhao *et al.*, 2022; Willi *et al.*, 2022). While LOLA focuses on opponent shaping, it does not explicitly address collaborative aspects of multi-agent settings. Social influence approaches encourage coordination and communication by rewarding agents for having causal influence over others' actions (Jaques *et al.*, 2019). Although these methods identify where an agent influences others, they do not involve explicit models of other agents' understanding or policies.

In all these approaches, the primary goal is to learn the assistant agent's policy, with other agents' policies represented implicitly. Typically, agents learn policies based on observable states and possibly the actions of others, without explicitly modeling the hidden goals of teammates. Interactive Partially Observable Markov Decision Process (I-POMDP) is a model that extends the standard POMDP framework to multi-agent settings by modeling other agents as part of the environment, explicitly representing their beliefs, intentions, and decision-making processes (Gmytrasiewicz and Doshi, 2005). This recursive reasoning enables agents to plan while accounting for the presence and potential strategies of others. However, while I-POMDPs provide a principled approach to modeling interactions, they can be computationally intractable and do not specifically target collaborative dynamics or goal alignment in fully cooperative scenarios. AssistanceZero is another recent work that explicitly formulates assistance as a two-player game and extends AlphaZero with a neural network that predicts human actions and rewards, enabling deep planning under uncertainty in environments with vast goal spaces (Laidlaw *et al.*, 2025). In contrast to these, our work emphasizes lightweight, cognitively inspired goal inference: we aim to reach similar reasoning processes as a toddler knowing to assist an adult, even if they have never been in a similar situation before, due to inherent altruistic motives Warneken and Tomasello (2006). This approach is based on observable behavior, avoiding the need for complex predictive models or retraining. This makes it more suitable for realistic, dynamic settings, such as human-robot collaboration, where goals may change over time and agents must adapt quickly with limited computation (Masters and Sardina, 2019; Shamir *et al.*, 2024; Shamir and Mirsky, 2025).

In this work, we focus on fully cooperative multi-agent settings, where all agents share a common goal (which may not be directly accessible to all teammates) and they work together to achieve it (Grosz and Kraus, 1999). Currently, we do not assume the existence of any explicit communication protocol between the agents; instead, coordination emerges implicitly through shared objectives and observed behavior. However, we recognize the potential benefits of incorporating communication and prefer representations that will allow us to enhance coordination and adaptability using communication in the future. To comply with this desiderata, we refer to the BDI literature and specifically, to Hidden Goal Markov Decision Processes (Fern *et al.*, 2014):

Hidden Goal Markov Decision Processes (HGMDPs) HGMDPs are specialized MDP-based models designed to formalize the problem of assistive agents aiding goal-directed users whose objectives are not directly observable.

Fern *et al.* (2014) defined a Hidden Goal Markov Decision Process (HGMDP) as a tuple $\mathcal{M} = (S, A, A', G, T, R, I, G_0, \pi)$, where S is the set of world states, A is the set of actions available to the leading agent (e.g., the user), and A' is the set of actions available to the assistant agent. G is a finite set of possible goals for the leading agent, where each goal $g \in G$ represents a set of desired world states such that $g \subseteq S$. The transition function $T : S \times (A \cup A') \times S \to [0, 1]$ defines the probability of transitioning between states given an action, and the reward function $R : S \times (A \cup A') \to \mathbb{R}$ assigns a real-valued cost to each action in a given state. $I : S \to [0, 1]$ is the initial state distribution, and $G_0 : G \to [0, 1]$ represents the prior distribution over the agent's goals. Finally, $\pi : S \times G \to \Delta(A)$ denotes the (unknown) policy of the leading agent, specifying a distribution over actions conditioned on the current state and goal.

Hidden Goal Markov Decision Processes (HGMDPs) introduce a belief state—a probability distribution over possible goals—that serves as a sufficient statistic for planning under uncertainty (Fern *et al.*, 2014). In this framework, the assistant observes the world state and the actions of the leading agent but does not have direct access to the agent's goal $g \in G$. The assistant's objective is to select actions from A' that assist the leading agent in achieving its goal, thereby minimizing the expected cumulative cost over an episode. The assistant must infer the agent's goal based on observed behaviors and select assistive actions accordingly. Note that in this setting, the leader is assumed to be non-learning. However, it may follow a stochastic policy or one that depends on the actions of the assistant. These assumptions are consistent with the ad hoc teamwork literature, which emphasizes spontaneous cooperation, an aspect closely aligned with the focus of our work (Mirsky *et al.*, 2022).

Figure 1 provides a running example of HGMDP using the hungry-thirsty environment (Singh *et al.*, 2009). The leader (depicted with a moustache) is either hungry or thirsty, and the assistant must help by fetching the appropriate item—sushi for hunger or water for thirst. A denotes the leader's actions, A' represents the assistant's actions, and G_0 is the assistant's initial belief about the leader's goal. A full description of this environment is provided in Section 4.

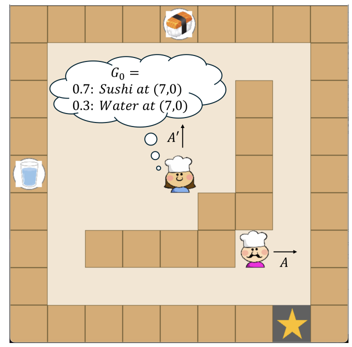

Figure 1: An illustration of HGMDP in the hungry-thirsty environment.

2.2 Goal Recognition with MDPs

The prior section discussed HGMDPs, where a critical challenge for the assistant agent is to understand the leading agent's goals. Several frameworks have been proposed to address goal recognition and observer-aware planning in the context of MDPs. Some of these representations are a single agent point-of-view where the observer is outside of the modeled world, while others are fully multi-agent in the sense that they model both the leader and the assistant.

- **Goal Recognition over POMDPs**: This approach involves inferring a probability distribution over possible goals of an agent whose behavior results from a POMDP model. The observer shares the POMDP model with the agent as common knowledge, except for the agent's true goal. The task is to compute the posterior goal distribution based on observed actions (Ramırez and Geffner, 2011).
- **Bayesian Delegation**: In this multi-agent settings, Bayesian Delegation enables agents to rapidly infer the hidden intentions of others by inverse planning, all share a similar model of the world. Agents coordinate their high-level plans and low-level actions without prior experience, demonstrating effective ad-hoc collaboration (Wu *et al.*, 2021).
- **Goal Recognition as Reinforcement Learning**: Amado *et al.* (2022) introduced a framework that combines model-free reinforcement learning and goal recognition. The approach involves offline learning of policies or utility functions for each potential goal and online inference to determine the most likely goal based on observations. This method alleviates the need for manual domain modeling and enables goal recognition in complex environments.
- **Observer-Aware MDPs (OAMDPs)**: OAMDPs provide a framework for producing observer-aware behaviors, where an agent considers the beliefs of an observer when planning its actions (Miura and Zilberstein, 2021). This framework aims to improve the interpretability of agent behaviors and is less complex than I-POMDPs (Gmytrasiewicz and Doshi, 2005).
- **Partially Observable Markov Chain of Plans (POMCoP)**: POMCoP is a system designed for planning in collaborative domains, where an AI sidekick assists a human player (Macindoe *et al.*, 2012). It operates by reasoning about how its actions will affect its understanding of humans' intentions, effectively maintaining a belief over possible human goals.

While all these frameworks address aspects of goal recognition and observer-aware planning with MDPs, which are MARL-compatible representations, we focus on HGMDPs rather than these other approaches as they offer a more comprehensive model by integrating goal inference and assistive action selection within a unified framework. HGMDPs are particularly well-suited for MARL scenarios where agents must cooperate without access to the leader's goal or knowledge of the world, as they allow for planning under uncertainty over different goals using belief states.

3 Translating a MARL Scenario into HGMDP

The HGMDP formalism provides a principled framework for modeling goal uncertainty in multi-agent settings where agents must collaborate without access to each other's internal states. By maintaining a *belief state*—a probabilistic estimate over the possible goals of the leading agent—the assistant (or follower) can plan and act under uncertainty, rather than committing to a single, fixed hypothesis of the leading agent's goal. This representation enables flexible and goal-aware decision-making that adapts to ambiguity and evolving evidence over time. For instance, consider a case where the assistant assigns equal probability ($50\% - 50\%$) to two goals that require opposing responses. If goals are treated as part of the observable state rather than as latent variables, the agent can at best learn a policy aligned with one goal, effectively ignoring the other. In contrast, reasoning over a belief distribution allows the agent to optimize its behavior over the whole spectrum of possible goal distributions, taking into consideration that the learned behavior only suits 50% of the agent's belief, thus enabling more robust and anticipatory assistance. In this way, the belief update mechanism serves as a critical bridge between low-level observations and high-level inference, supporting more adaptive and intelligent cooperative behavior.

To illustrate this framework in a practical setting, we consider a simplified leader-assistant scenario inspired by the **Hungry-Thirsty** domain (Singh *et al.*, 2009). In this two-agent system, the leader has one of two latent goals—either reaching a food cell (if hungry) or a water cell (if thirsty). The assistant's task is to assist the leader in reaching that goal as efficiently as possible. The assistant does not receive explicit communication of the leader's internal state or goal. Instead, it must infer the leader's goal solely on observed behavior and environmental state.

Case Study: Incorporating Goal Recognition using HGMDP in Hungry-Thirsty

In this domain, the state space includes the positions of both agents and the locations of food and water, while the goal space G consists of two elements: `hungry` and `thirsty`. The leader's policy is conditioned on its goal, which is hidden from the assistant. We model the leader's goal as hidden, transforming the problem into an HGMDP. The assistant maintains a belief over the leader's goal, denoted $b_t(g) = P(g \mid h_t)$, where h_t is the interaction history up to time t.

This belief is then used as an input to the assistant's policy, $\pi' : (s_t, b_t) \to A'$, enabling it to adapt its actions based on the inferred goal. For instance, if the leader appears to be heading toward the food location, the assistant may infer that the leader is hungry and either assist in retrieving the food or allocate effort elsewhere if the leader is already near the target. In this way, the assistant exhibits rudimentary theory of mind—reasoning about not only *what* the leader is doing but also *why*.

This formulation offers several advantages:

- **Modularity:** The belief update and policy learning processes can be decoupled, allowing for independent improvements and more interpretable agent behavior.
- **Efficiency:** By focusing on likely goals, the assistant avoids learning to assist in irrelevant or low-probability scenarios.
- **Interpretability:** Goal inference provides a transparent rationale for assistive behavior and facilitates debugging and trust.

We propose that this structure of combining MARL, HGMDPs, and goal recognition serves as a promising baseline for future work, including leveraging various RL algorithms for training, as well as tackling more complex cooperative domains such as OVERCOOKED with longer, more complex recipes. While the Hungry-Thirsty domain is minimal, it captures the core challenge of latent-goal inference and provides a foundation for extending to richer scenarios.

Modeling Assumptions and Design Choices

To simplify implementation while maintaining generality, we make the following assumptions:

- The leader's policy is goal-directed and stochastic but does not explicitly model the assistant.
- The assistant observes the environment and the leader's actions, but not the leader's internal state.
- The goal prior distribution G_0 is known or can be estimated from offline data.
- The environment dynamics are either deterministic or known, enabling tractable belief updates.

4 Experimental Setup

Environment and Agents We evaluate the assistant agent in the "Hungry-Thirsty" environment, a grid-based simulation where agents navigate walkable tiles while avoiding static obstacles (counters). Experiments were conducted on three distinct layouts (Figure 2), with fixed starting positions for the leader and assistant agents.

Trial Configuration For each layout, we ran 50 trials. In each, 2–4 food items (sushi, water, egg, bread) were randomly placed on counters, and a target item for the leader was selected from among them. This setup models a leader-assistant scenario where the assistant observes the leader and attempts to assist.

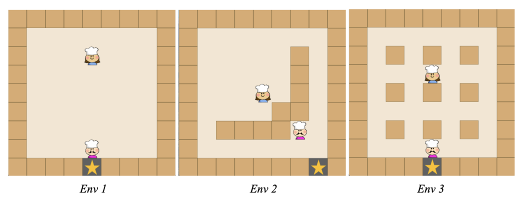

Figure 2: The Hungry-Thirsty environments used in experiments

Agent Behavior The leader begins each trial in "independent" mode, following an optimal, deterministic path to its goal. The assistant observes the leader's movements to infer its goal by tracking changes in shortest path distances from the leader's starting position to each food item. A greater decrease to a specific item suggests the leader is targeting it. Upon inference, the assistant fetches and delivers the item, then returns to its start location. This reflects a one-task model; future work will consider sequential tasks. Once the assistant acts, the leader switches to Bayesian Delegation (Wu *et al.*, 2021) to coordinate. Trials end when the leader reaches the delivery station with the target item.

Trial Selection Criteria Only trials where the assistant can unambiguously infer the leader's goal (before item pickup) are included. This ensures 100% goal recognition accuracy, allowing analysis to focus on:

1. Timesteps required for goal recognition
2. Assistant's success in fetching and delivering the item (not always 100%)

Baseline Approach The setup does not introduce new algorithms. The leader follows optimal pathfinding and later uses Bayesian Delegation; the assistant uses deterministic inference and optimal navigation. This baseline measures idealized agent performance to inform future research on more complex, adaptive behaviors.

Research Questions In this work, we explore:

- Does more goal ambiguity (more items) increase recognition time?
- Does goal ambiguity reduce delivery success rate?
- Is earlier goal recognition correlated with successful delivery?

- Just how useful is the assistant to the leader? How much time does the assistant save the leader, on average, during task completion?

This design evaluates a baseline assistant using simple inference and navigation in a human-robot coordination task, providing a foundation for future work on adaptive, learning-based agents.

Metrics and Evaluation As we focus this preliminary investigation on setting up a clear baseline for evaluation of a cooperation between a leader and an assistant agent, we track:

- **Steps to Goal Recognition**: Average timesteps until correct inference
- **Success Rate (Delivery)**: Whether the assistant delivered the correct item
- **Steps to Goal Recognition (Success Only)**: Same as above, but only for successful deliveries
- **Average Contribution of the Assistant**: The average reduction in steps taken/required (not sure which of these two words I should use here) by the leader when the assistant is present. We assume the leader stops pursuing its goal as soon as the assistant makes their first movement. The assistant's first movement indicates that they have recognized the goal and begun fetching it, so the leader no longer needs to take action.
- **Average Contribution of the Assistant (Success Only)**: Same as above, but only for successful deliveries.

5 Preliminary Results

Our experiments revealed several key trends in how the assistant agent's performance varied as the number of potential goals (food items) increased, illustrating the strengths and limitations of simple goal inference within our controlled evaluation setting.

Time to Goal Recognition Increases with More Goals The timesteps needed for the assistant agent to recognize the leader's goal increased when there were more food items in the environment. As shown in Table 1, the average number of timesteps until recognition increased across all three environments as the number of potential goals increased from two to four. In Environment 1, the average number of timesteps until recognition increased from 1.88 timesteps with two goals to 4.46 timesteps with four goals. Similar trends were observed in Environment 2 (2.77 to 6.30 timesteps) and Environment 3 (2.10 to 4.31 timesteps). This pattern suggests that when more food items (potential goals) are present in the environment, the assistant agent must observe more movement from the leader before it can disambiguate and confidently determine the target item.

Env.	2 Goals	3 Goals	4 Goals
1	1.88	2.70	4.46
2	2.77	4.22	6.30
3	2.10	2.86	4.31

Table 1: Timesteps to Goal Recognition

Env.	2 Goals	3 Goals	4 Goals
1	88%	45%	38%
2	50%	44%	30%
3	40%	29%	25%

Table 2: Success Rate (Delivery)

Env.	2 Goals	3 Goals	4 Goals
1	1.87	2.22	2.40
2	1.36	1.88	2.67
3	1.38	1.25	1.75

Table 3: Timesteps to Successful Delivery

Success Rate (Delivery) Decreases with More Goals As a consequence of the increased time required for goal recognition when there are more potential goals, the assistant agent's overall success in fetching and delivering the target food item **declined** with more goals. The data for successful fetch and delivery in Table 2 illustrates this. Environment 1 saw a decrease from 88% success with two goals to 38% with four goals. Similar trends were observed in Environment 2 (50% to 30%) and Environment 3 (40% to 25%). These results highlight that not being able to recognize the leader's goal early on makes it highly difficult for the assistant to fetch and deliver the item before the leader reaches the item on their own.

Earlier Recognition Correlates with Successful Delivery When the assistant agent was able to successfully complete the fetch and delivery, it tended to recognize the goal earlier in the trial. This is supported by Table 3, which shows the average time until recognition when there is a successful fetch and delivery. Comparing Table 3 to Table 1, we can see that for each environment and number of goals, the average time to recognition when there is a successful fetch and delivery is always less than the overall average time until recognition shown in Table 1. This indicates that faster identification of the target item is associated with a higher likelihood of a successful delivery.

Assistant Can Significantly Reduce Leader Steps, But Contribution Diminishes with More Potential Goals The assistant saves the leader significant time, even when trials with un-successful delivery are factored in. Table 4 shows that across all environments and numbers of potential goals, the assistant reduces the number of steps taken by the leader by at least 14%. For trials with three goals or fewer, this number jumps to 20%. As expected, the assistant's positive contribution decreases as the number of potential goals increases. This is because, as shown in Table 2, the assistant's success rate drops with more potential goals, leading to more trials where the leader has to retrieve the item independently. In these specific trials, the leader experiences no step reduction.

Env.	2 Goals	3 Goals	4 Goals
1	59.9%	27.8%	24.7%
2	37.5%	28.9%	19.6%
3	26.6%	20.9%	14.5%

Table 4: Average Contribution of the Assistant

Env.	2 Goals	3 Goals	4 Goals
1	67.9%	61.8%	64.2%
2	75.0%	65.1%	65.4%
3	66.6%	73.0%	58.0%

Table 5: Average Contribution (Success Only)

Major Step Reduction in Trials With Successful Delivery Based on Table 5, when the assistant successfully fetches and delivers the target item, it consistently leads to a significant step reduction for the leader, averaging approximately 65% across all environments and numbers of potential goals.

6 Conclusion

In this preliminary work, we explored a Theory of Mind (ToM)-inspired approach to enhance spontaneous cooperation in Multi-Agent Reinforcement Learning (MARL) settings. We drew inspiration from the Hidden Goal Markov Decision Process (HGMDP) framework to model the interaction between a goal-driven agent and an assisting agent, in order to reduce the cognitive and computational complexity required for effective assistance (Amado *et al.*, 2022; Fern *et al.*, 2014). Our investigation focused on enabling one agent to recognize and assist the goal of another without extensive reasoning or internal simulation (Masters and Sardina, 2019; Shamir *et al.*, 2024).

We translated a MARL scenario into an HGMDP framework using a simplified leader-assistant "Hungry-Thirsty" domain (Wu *et al.*, 2021). In this setup, the assistant agent observed the leader's movements to infer its goal based on changes in the shortest walkable path to potential target items. This deterministic approach to goal recognition and optimal pathfinding for navigation served as a baseline to evaluate idealized agent performance.

Our preliminary results indicated several key trends. Firstly, the time required for the assistant agent to recognize the leader's goal increased as the number of potential goals (food items) in the environment grew. Consequently, the overall success rate of the assistant in fetching and delivering the target item declined with a higher number of potential goals, as delayed recognition made it more difficult for the assistant to intervene effectively before the leader reached the item independently. Furthermore, our findings showed a correlation between earlier goal recognition and a higher likelihood of successful fetch and delivery by the assistant agent.

This study represents an initial step, and the presented system has limitations, including the use of a simplified environment and a deterministic, non-learning assistant. However, the framework of combining MARL with HGMDPs for goal recognition offers a promising foundation for developing more sophisticated assisting agents. Future work will focus on enhancing the reasoning and execution capabilities of assisting agents, potentially by leveraging various reinforcement learning algorithms for training. We also plan to extend this approach to more complex cooperative domains that involve longer and more intricate tasks. Ultimately, this line of research aims to create agents that can intuitively and effectively collaborate in dynamic multi-agent systems.

References

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

Leonardo Amado, Reuth Mirsky, and Felipe Meneguzzi. Goal recognition as reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 9644–9651, 2022.

Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, March 2003.

Alan Fern, Sriraam Natarajan, Kevin Judah, and Prasad Tadepalli. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research*, 50:71–104, 2014.

Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 122–130, 2018.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *Intelligent Workshop on Agents Theories, Architectures, and Languages (ATAL)*, pages 1–10, 1999.

Piotr J. Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

Barbara J Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

Barbara J Grosz and Sarit Kraus. The evolution of sharedplans. In *Foundations of rational agency*, pages 227–262. Springer, 1999.

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 *International Conference on Machine Learning (ICML)*, pages 3040–3049, 2019.

Cassidy Laidlaw, Eli Bronstein, Timothy Guo, Dylan Feng, Lukas Berglund, Justin Svegliato, Stuart Russell, and Anca Dragan. AssistanceZero: Scalably solving assistance games. *arXiv preprint arXiv:2504.07091*, 2025.

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in Artificial Intelligence*, 5:778852, 2022.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Neural Information Processing Systems (NeurIPS)*, 2017.

Owen Macindoe, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. POMCoP: Belief space planning for sidekicks in cooperative games. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, pages 38–43, 2012.

Peta Masters and Sebastian Sardina. Cost-based goal recognition in navigational domains. *Journal of Artificial Intelligence Research*, 64:197–242, 2019.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. A survey of ad hoc teamwork: Definitions, methods, and open problems. In *European Conference on Multiagent Systems (EUMAS)*, pages 1–8, 2022.

Shuwa Miura and Shlomo Zilberstein. A unifying framework for observer-aware planning and its complexity. In *Uncertainty in Artificial Intelligence (UAI)*, pages 610–620, 2021.

Miquel Ramırez and Hector Geffner. Goal recognition over POMDPs: Inferring the intention of a POMDP agent. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2009–2014, 2011.

Anand S Rao, Michael P Georgeff, et al. Bdi agents: from theory to practice. In *Icmas*, volume 95, pages 312–319, 1995.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

João G Ribeiro, Cassandro Martinho, Alberto Sardinha, and Francisco S Melo. Assisting unknown teammates in unknown tasks: Ad hoc teamwork under partial observability. *arXiv preprint arXiv:2201.03538*, 2022.

Matan Shamir and Reuth Mirsky. Graml: Goal recognition as metric learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Matan Shamir, Osher Elhadad, Matthew E Taylor, and Reuth Mirsky. ODGR: Online dynamic goal recognition. *arXiv preprint arXiv:2407.16220*, 2024.

Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from. In *Annual Conference of the Cognitive Science Society*, pages 2601–2606, 2009.

Ming Tan. Multi-agent reinforcement learning: Independent versus cooperative agents. In *International Conference on Machine Learning (LCML)*, page 330–337, 1993.

Felix Warneken and Michael Tomasello. Altruistic helping in human infants and young chimpanzees. *science*, 311(5765):1301–1303, 2006.

Timon Willi, Alistair Hp Letcher, Johannes Treutlein, and Jakob Foerster. COLA: consistent learning with opponent-learning awareness. In *International Conference on Machine Learning (ICML)*, pages 23804–23831, 2022.

Sarah A Wu, Rose E Wang, James A Evans, Joshua B Tenenbaum, David C Parkes, and Max Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13:414–432, 2021.

Stephen Zhao, Chris Lu, Roger B Grosse, and Jakob Foerster. Proximal learning with opponent-learning awareness. In *Neural Information Processing Systems (NeurIPS)*, pages 26324–26336, 2022.