Title

Imagined We: Understanding and Bridging the Gap Between Human Cooperation and Multi-agent Reinforcement Learning

Author

Zhao, Minglu

Publication Date

2023

UNIVERSITY OF CALIFORNIA

Los Angeles

Imagined We: Understanding and Bridging the Gap
Between Human Cooperation and Multi-agent Reinforcement Learning

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Minglu Zhao

2023

ABSTRACT OF THE THESIS

Imagined We: Understanding and Bridging the Gap Between Human Cooperation and Multi-agent Reinforcement Learning

by

Minglu Zhao
Master of Science in Statistics
University of California, Los Angeles, 2023
Professor Tao Gao, Chair

Cooperation is a fundamental characteristic of humans that enables complex social interactions and joint achievements beyond the capacity of individuals. Multi-agent reinforcement learning (MARL) is one prevailing approach employed to model such cooperative behavior. While MARL as a generic algorithm serves well for modeling agents' different social intentions through adjusting the relations between their reward functions, theories from cognitive science suggest that MARL per se does not suffice for modeling human-unique cooperation. During cooperation, humans spontaneously establish a sophisticated framework of shared agency, in which a joint representation of an imagined central agent emerges automatically and owns normative power. We delve into the discrepancies between these two cooperation paradigms through two case studies. Inspired by the theory of shared intentionality in cognitive science, we further introduce the *Imagined We* framework, a novel approach that emulates human behavior across various tasks requiring joint efforts and communication.

The thesis of Minglu Zhao is approved.

Hongjing Lu

Ying Nian Wu

Tao Gao, Committee Chair

University of California, Los Angeles

2023

To my parents for all the love and support

To all those who have inspired and encouraged me

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

Cooperation is an essential characteristic of human multi-agent interaction systems that enables agents to achieve what individuals cannot accomplish alone. Anthropological studies suggest that cooperation is indeed a human-unique behavior where we form complex social institutes with the most advanced mode of coordination [BR09]. Modern advancements in Artificial Intelligence (AI) have been tailored towards simulating efficient multi-agent interactions across a myriad of tasks, generating intricate behavior within gaming platforms as well as real-life domains. One major area of machine learning focusing on behavior modeling is reinforcement learning (RL). RL is a prominent learning-based model with deep roots in psychology and neuroscience [LMB21, PBA21]. Through trial-and-error, RL approximates the optimal action policy achieved by maximizing long-term expected rewards [SB18]. Multi-agent reinforcement learning (MARL) extends RL to multi-agent settings and has been successfully applied to various challenging group coordination scenarios, such as autonomous-driving coordination [SSS16], as well as teaming in games like Dota 2 [BBC19] and StarCraft [VBC19].

However, while MARL has demonstrated high-quality coordination that parallels human behavior, fundamental disparities exist between the MARL paradigm and human cooperation as viewed from cognitive and social science perspectives. MARL approaches generally define cooperation as having all agents maximize a joint reward [BBD08]. This perspective pre-determines cooperation before the learning phase begins, positioning joint efforts on a task as the singular objective. Consequently, team members have neither the incentive to deviate from collaboration nor any alternative but to cooperate in order to optimize team reward. With the pre-assumed incentive to cooperate, the only concern remaining for MARL is how to cooperate in the best possible way. Contrarily, from a social and cognitive science lens, real-world cooperation diverges significantly from this setup, extending beyond agents passively working on a group task. While agents working together as a team want to maximize the group reward, in reality there also exists subsidiary components of agents' reward functions that originate from the self interests of individuals, minimizing action costs, for example. The existence of such individualism mandates agents to maintain a separate set of individual reward objectives, the accomplishment of which might necessitate actions that are against the cooperative goal. Despite these complexities, cognitive theories posit that humans still display cooperative behavior even from an early age, likely due to humans' unique ability to represent teammates as equal partners forming a shared agency [Tom19]. Human cooperation thus morphs into a mixed-interest problem that requires the harmonious coordination of individual interests. Mixed-interest tasks like prisoners' dilemma and tragedy of the commons are also treated as the most fundamental questions for study in economics and sociology. With this reasoning, it is indeed unclear whether the successful coordination behavior modeled by AI algorithms are truly cooperative from a human-cooperation perspective, where the potential challenges in cooperation have been avoided in the problem setting by default. To build AI models that can truly mimic human cooperation and thus more efficiently serve as humans' partners in cooperative tasks, efforts in modeling human-unique cognitive processes of cooperation is indispensable.

With the inherent differences between human cooperative behavior and the approach employed by MARL algorithms, we design a series of modeling experiments to identify the potential gaps that hinders modeling truly human-like cooperation. Inspired by the theory of shared intentionality in cognitive science, we further introduce a new framework for cooperation modeling called *Imagined We*, which successfully emulates human behavior across various tasks that require joint efforts as well as communication.

The thesis is organized as follows: Chapter 2 provides a background in both MARL and cognitive science, detailing their respective approaches to cooperation. I first review the state-of-the-art approaches in MARL focusing on modeling cooperative behavior; then I review the major cognitive science perspectives on how humans, among all creatures, uniquely cooperate with each other, with a focus on highlighting the differences with current AI modeling perspectives. Chapter 3 presents a series of modeling experiments for examining the potential gap between MARL and human-like cooperation. In Chapter 4, I define the *Imagined We* framework and explain the results of our experiments. Finally in Chapter 5, I suggest potential future research directions as part of our ongoing work.

CHAPTER 2

Background and Related Work

In this chapter, I describe the approaches towards modeling cooperation by the MARL community, as well as the cognitive science perspective towards cooperation, aiming to draw a comparison between the two to lay the foundation for our methodology.

2.1 Multi-agent reinforcement learning for cooperation

2.1.1 Reinforcement learning and Markov decision process (MDP)

Reinforcement learning is an area of machine learning where the model learns how to act, which is a mapping from environment to actions, to maximize a numerical reward signal [SB18]. While no direct supervision is required, RL learns by directly interacting with the environment through trial-and-error. From the perspective of dynamical systems, RL problems can be framed as Markov Decision Processes (MDP), where a decision maker interacts with the environment. At each timestep t, the agent selects action $a_t \in A$ based on the environment state $s_t \in S$. The action would lead to some reward in the next timestep $r_{t+1} \in R$ by reward function $r(s, a) = \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a]$ and transits to a new state s_{t+1} by transition function $P(S_t = s' \mid S_{t-1} = s, A_{t-1} = a)$. The goal of reinforcement learning is to maximize the expected discounted cumulative sum of future rewards $\mathbb{E}[\sum_{t=0}^{\infty} \gamma r_t]$, where γ is a discount factor.

Various frameworks have been proposed to achieve this objective. One major branch of work is the value-based reinforcement learning, where the goal of training is to find the optimal policy value function $Q^{\pi}(s,a) = \mathbb{E}_{\pi}[R_t \mid S_t = s, A_t = a]$. The agent then acts by taking the action that maximizes the Q function using the idea of Bellman update:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)) \tag{2.1}$$

With the development of Deep Learning, neural networks structures have been introduced to RL frameworks. In the value-based RL paradigm, value networks are designed as deep networks to allow for application to higher-dimensional state spaces, with parameters trained through gradient descent to minimize Q value prediction error [MKS15, HS15]:

$$\theta_{k+1} \leftarrow \theta_k - \alpha \nabla_\theta \mathbb{E}_{s' \sim P(s'|s,a)} \left[\left(Q_\theta(s,a) - (r_{t+1} + \gamma \max_{a'} (Q_{\theta^-}(s', a') \right)^2 \right] \Bigg|_{\theta = \theta_k}, \tag{2.2}$$

where Q_{θ^-} is a target network that is a delayed copy of the training network, usually incorporated to stabilize training.

Another line of work aims to directly improve policies $\pi_\theta(a \mid s) = P(a \mid s)$ using the policy gradient theorem, which provides a gradient ascent method for updating policy parameters θ to maximize the reward (Equation 2.3). With such parameterization, the policy directly generates the action to take, and thus the framework can be applied to environments with continuous action spaces as well as to generate stochastic policies.

$$\begin{aligned} J(\theta) &= E_{\pi_\theta}[\sum_{t=0}^{\infty} \gamma r_t] \\ \nabla_\theta J(\theta) &= E_\pi[\nabla_\theta(\log \pi(a \mid s))R(s,a)] \\ \theta &\leftarrow \theta + \alpha \nabla_\theta J(\theta) \end{aligned} \tag{2.3}$$

Combining the two frameworks yields another popular framework named actor-critic methods, where an actor network generates actions evaluated by a critic network. This method capitalizes on the strengths of both strategies and is utilized in various state-of-the-art algorithms [LHP15, SLA15, SWD17, HZA18].

2.1.2 Multi-agent reinforcement learning

Multi-agent RL (MARL) builds upon the single-agent RL framework and is applied to scenarios with more than one agent, where the environment dynamics are influenced by the joint actions of all agents [Lit94]. Under the Multi-agent Markov Decision Process (MMDP) framework, the goal of training is to find a joint policy set $\pi = (\pi_1, ..., \pi_n)$ to maximize the expected return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma r_t]$, where each agent $i = 1...n$ has one policy $\pi_i(s) : S \rightarrow A$.

One way of transitioning from single-agent to multi-agent settings is to model each agent independently, treating others as part of the environment [Tan93, TMK17]. This decentralized model is scalable and suits partially observable scenarios but can face stability and convergence issues due to the non-stationarity of the environment as agents' policies evolve – changes in the environment are not explainable by the change of the agent's own policy [MLL12]. This limitation has been recognized in recent works with various stabilization techniques proposed under the decentralized framework [OPA17,FNF17,ZYL18,SYZ20]. Alternatively, a centralized training approach considers all agents' observations and actions, producing state-action values reflecting their coordination's effectiveness [HBZ04]. Here, the environment is seen as stationary, even when agents' policies change, avoiding the non-stationarity issue. However, scalability becomes a challenge as the number of agents increases.

Modern MARL approaches often fall between these two design extremes and train agents using the paradigm of centralized training with decentralized execution (CTDE) [HBZ04, OSV08,SLG17,RSS18]. Actor-critic training [SB18] is well adapted to this framework, where decentralized actors decide the action policy only through local information, and centralized critics evaluate the group performance using augmented information, including other agents' policies and potential communication [LWT17, FFA18, IS19]. The centralization exists only in evaluation but not in control, since otherwise the algorithm may be considered as a single-agent learning problem. Critically, the framework can be applied to scenarios with different social intentions in a similar way by adjusting the relations between agents' reward functions [LWT17]. For competition, the reward functions are zero-sum; for cooperation, agents align their rewards through the same reward function [BBD08].

Modern MARL strategies often navigate the middle ground between the two design extremes, utilizing a paradigm known as centralized training with decentralized execution (CTDE) [HBZ04, OSV08, SLG17, RSS18]. The actor-critic training technique fits well within this framework [SB18]. Here, decentralized actors use local information to form action policies, while centralized critics assess the group's performance using expanded information, including other agents' policies and potential communication [LWT17, FFA18, IS19]. Critically, centralization is reserved for evaluation, not control, otherwise, the algorithm would essentially become a single-agent learning problem. Conveniently, MARL offers a generic solution to different cooperative [SZL21, ZYL18], competitive [SSS17, XCW20, ZKB20], or mixed interest settings [LP02] by modulating the relationship between agents' reward functions. Competitive settings are often treated as zero-sum games in which the reward gained by one agent is exactly the loss of the other. When the environment is fully cooperative, all n agents typically observe the same joint reward value r_t at each time step [BBD08], effectively sharing the reward evenly among group members. The focus of our work is primarily on MARL modeling in cooperative environments, exploring how such a setting impacts the agents' coordination behavior.

2.2 Cognitive science perspectives towards cooperation

2.2.1 Challenges faced by MARL in mixed interests scenarios

MARL enjoys the flexibility of being able to train agents with different social intentions through adjusting the relations between reward functions. The problem setup intuitively suggests that as long as a reward function is properly specified, agents should be able to learn intelligent behavior of any kind [SSP21]. However, when facing general-sum games that are beyond pure cooperation or competition, efficient solutions are far from guaranteed using the current MARL approaches [SC96,LZL17]. One classic prototype of the multi-agent cooperation challenges faced by people is the Stag Hunt problem [Sky04]. In Stag Hunt, two hunters must decide individually whether to hunt a stag or a hare. Only when the two both hunt for the stag can they obtain the stag reward, which is higher than the reward of the hare. On the other hand, hunting alone for the hare renders minimal risk. In other words, the two agents choose between whether to ambitiously hunt the stag and expect the partner to think the same way while taking the risk, or to give up on cooperation and the higher benefits and safely hunt the hare. From a game theory perspective, there are two pure-strategy Nash equilibria: one that both agents cooperate and the other that both defect. Without a cooperative mindset, agents directly trained with the reward function matrix may easily converge to the local minima where both defect.

A similar idea lies in the formation of people's bargaining behavior. Members of a group usually hold distinct skill sets and reward structures. While everyone is aware that jointly working on a task facilitates its completion, there is little guarantee that individual members will contribute to the group activity as expected. From a reward-maximization perspective as suggested by MARL approaches, coordination in such cases might not be the optimal solution. Imagine a scenario with a farmer and a baker: the baker needs wheat to make bread, while the farmer can produce wheat but needs bread to survive. Intuitively, the best case scenario would be for both parties to cooperate by forming a trading system. However, since both farming and baking takes time and effort, both parties may encounter more loss if the coordination system is not guaranteed. To avoid potential threats from the other party defecting, as in the stag hunt game, the reward-maximization agents may simply abandon the joint reasoning and only work for self interests by picking the safe option.

Towards the extreme of such self-interest behavior further leads to the famous phenomenon of tragedy of the commons: individuals with access to a repertoire of limited public goods act by their self-interest, which causes the depletion of the resources [Har68].

Since the agents are purely driven by the reward-maximization principle, taking all resources whenever possible is the optimal way from the individual perspective. Without considering group benefits, group members' behavior are free of moral constraints including ownership and fairness which guide prosocial behaviors. Intuitively, such a situation hampers the future coordination of the group. In order for cooperation to emerge from the scenario, agent' behavior should be effectively regulated with a collaborative mindset. Modeling this concept thus demands a formulation of the group perspective that is beyond acting to maximize rewards as proposed by MARL methods.

2.2.2 The cognitive science perspectives on how humans cooperate

On the other hand, while facing such mixed-interests dilemmas in everyday life, humans still manage to establish robust cooperation even when game theory suggests otherwise, such as when different groups have unmatched power and when repetitive interactions cannot be guaranteed. One case study on the producer-middleman relationships in precolonial Africa indicates that humans indeed reached a form of cooperative agreement by conforming to an imagined regulatory system [Lee14]. Under the potential threats posed by an imbalanced power between the producers and the middlemen, the game theory will suggest a result of renouncing cooperation from both sides as a local minimum Nash equilibrium solution. Nevertheless, humans choose to agree on an imagined credit system that enforces cooperation, creating an opportunity to depart from the suboptimal solution and achieving a greater collective reward as a result. Such evidence manifests the normative power and creativity of human cooperation that can emerge beyond trial-and-error training.

Indeed, cognitive studies indicate that effectively regulated human cooperation emerges from a young age and is further considered as human-unique by evolutionary studies comparing human behavior with chimpanzees [Tom19]. One signature characteristic of human cooperation is the enforcement of commitment. Developmental studies conducted with children indicate that humans starting from 2-3 years old demonstrate a strong tendency to maintain and regulate the cooperative structure of a group that agents should jointly commit to a task. In a group task where children prematurely receive their own share of the rewards, they still persisted in participating in the joint activity so that both partners ended up with rewards — and more often than if the partner just asked for help in a similar situation but outside of any collaboration [HWT12]. Further, 18-month-old toddlers attempt to re-engage partners when the partner leaves the joint activity [WCT06]. Young children jointly commit to tasks and often explicitly acknowledge this (e.g., A: "Let's X."; B:"OK") [MGK16]; When children break a commitment, they acknowledge their leaving [GBC09] and express guilt for doing so [VCT16]. In this case, for humans the notion of cooperation persists even independently of rewards, which serves as a social contract to which they expect both partners to commit.

To maintain and further promote cooperative behavior, children tend to view collaborators as equal partners jointly working on the task with respective roles. When roles are not adequately played by collaborative partners, children usually protest with normative words [KST18]. Developmental studies also demonstrate children's preference for fairness in collaborative tasks where they recognize teammates' efforts in a fair way. Under unequal distribution of spoils, 3-year-old children who got more spoils shared with their partner to equalize, which they did not do outside of collaboration [HWG11]. When a collaborative partner does not share the spoils fairly, children tend to protest against the behavior [WLM11]. With a strong conceptualization of fairness, young children further excluded free riders (as opposed to partners) from the spoils [MFT15].

In this way, starting early in development, humans form specific understandings of how cooperation should be carried out and what behavior should be regularized. Ideas from philosophy argues that this is likely due to humans' unique socio-cognitive ability to represent the self and others as a collective whole and intend the group to do something together, a concept referred to as the shared intentionality framework [Gil92, Bra92]. Under this framework, agents approach a group task as a whole, taking a joint perspective when planning for actions. This is essentially constructing a supraindividual agent “We” in one’s mind and asking oneself “what does ’We’ expect me and others to do?” This representation of the “We” intention is hypothesized to be the fundamental infrastructure differentiating humans from other animals [Tom19]. The shared intentionality framework thus serves as a stronger cooperation motivation with normative constraints that is beyond sharing rewards as defined in MARL approaches.

CHAPTER 3

Understanding the Gap

Inspired by the differences in interpretations between MARL and cognitive science studies regarding cooperation, we were curious about whether MARL models would be able to handle complex social dilemmas as faced by humans, using coordinated hunting as a case study. This chapter introduces two case studies we performed to better understand the coordination achieved by MARL.

3.1 General method

3.1.1 Test environment

Coordinated hunting is commonly observed in many species [MTS14, HSB97, GCE05, YY10, Bed88] and requires sophisticated team planning to execute efficiently. To analyze the performance of MARL in coordination tasks, in our experiments we adapt a previously developed non-cooperative hunting task for use in a cooperative environment [GNS09]. This task lies at the border between proposed evolutionary demands for cooperation and empirical studies of the same, exploring a current gap between the two. It has also been widely adopted in MARL research as a benchmark test [LWT17]. In the task, predator and prey agents are trained so that predators aim to catch the prey as much as possible, while prey aim to avoid the predator.

3.1.2 Algorithm

In order to comprehend the effectiveness of state-of-the-art MARL algorithms, we employ one of the most significant MARL algorithms, the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [LWT17], as a representation of the general methodology adopted in the field. This algorithm builds upon the classic actor-critic framework in reinforcement learning [SB18]. Here, actor networks (policy networks) generate actions, denoted as $\pi_{\theta^a}(a|s)$, using the network parameters θ^a. Critic networks (value networks) evaluate actions through action-value functions $Q_{\theta^c}(s,a)$ using the network parameters θ^c.

During training, the critic network minimizes the estimation error $L(\theta^c)$ between the network output $Q_{\theta^c}(s,a)$ and the target of update y:

$$L(\theta^c) = \mathbb{E}_{s,a,r,s'}\left[(Q_{\theta^c}(s,a) - y)^2\right]. \tag{3.1}$$

On the other hand, the actor network aims to maximize the objective function $J(\theta^a)$ by updating the network parameters following the policy gradient $\nabla_{\theta^a}J(\theta^a)$:

$$\nabla_{\theta^a}J(\theta^a) = \mathbb{E}_{s\sim p^\pi, a\sim\pi_{\theta^a}}\left[\nabla_{\theta^a}\log\pi_{\theta^a}(a|s)Q_{\theta^c}(s,a)\right]. \tag{3.2}$$

MADDPG is an extension of the actor-critic framework for multi-agent settings where each agent has its own actor and critic networks, allowing training of both competitive and cooperative agents. Detailed pseudocode for the MADDPG algorithm is presented in Algorithm 1 (Figure 3.1).

MADDPG has successfully demonstrated its capability in multi-agent coordinated hunting games where rewards are shared amongst agents [LWT17]. The training regimen of MADDPG exhibits two aspects of cognitive realism. Firstly, it employs cognitive constraints, meaning that each agent makes decisions based on its own observations without accessing information from other agents. Secondly, the model operates with cognitive intelligence,

Algorithm 1 Multi-Agent Deep Deterministic Policy Gradient (MADDPG)

```
for episode = 1 to M do
    Initialize random process $K$ for action noise
    Reset initial state $x$
    for timestep $t = 1$ to max-episode-length do
        for agent $i = 1$ to N do
            Make observation $o_i = O_i(x)$
            Take action $a_i = \mu_{\theta_i^a}(o_i) + K_t$
        end for
        Transit to new state $x'$ and obtain reward $r = (r_1, \ldots, r_N)$ based on $a = (a_1, \ldots, a_N)$
        for agent $i = 1$ to N do
            Make observation $o'_i = O_i(x')$
            Store $(o, a, r, o')$ to replay buffer $D$, $o = (o_1, \ldots, o_N)$, $o' = (o'_1, \ldots, o'_N)$
            Randomly sample a minibatch of $S$ samples $(o^j, a^j, r^j, o'^j)$ from $D$
            Set $y^j = r_i^j + \gamma Q_{\theta_i'^c}(o'^j, a'^j)$, where $a'_k = \mu_{\theta_k'^a}(o_j'^k)$
            Update critic by minimizing the loss:
            $L(\theta_i^c) = \frac{1}{S} \sum_j (y^j - Q_{\theta_i^c}(o^j, a^j))^2$
            Update actor by policy gradient:
            $\nabla_{\theta_i^a} J \approx \frac{1}{S} \sum_j \nabla_{\theta_i^a} \mu_{\theta_i^a}(o_i^j) \nabla_{a_i} Q_{\theta_i^c}(o^j, a_1^j, \ldots, a_i, a_N^j)$, $a_i = \mu_{\theta_i^a}(o_i^j)$
        end for
        Update target network parameters for each agent $i$:
        $\theta_i'^c = \tau \theta_i^c + (1 - \tau) \theta_i'^c$
        $\theta_i'^a = \tau \theta_i^a + (1 - \tau) \theta_i'^a$
    end for
end for
```

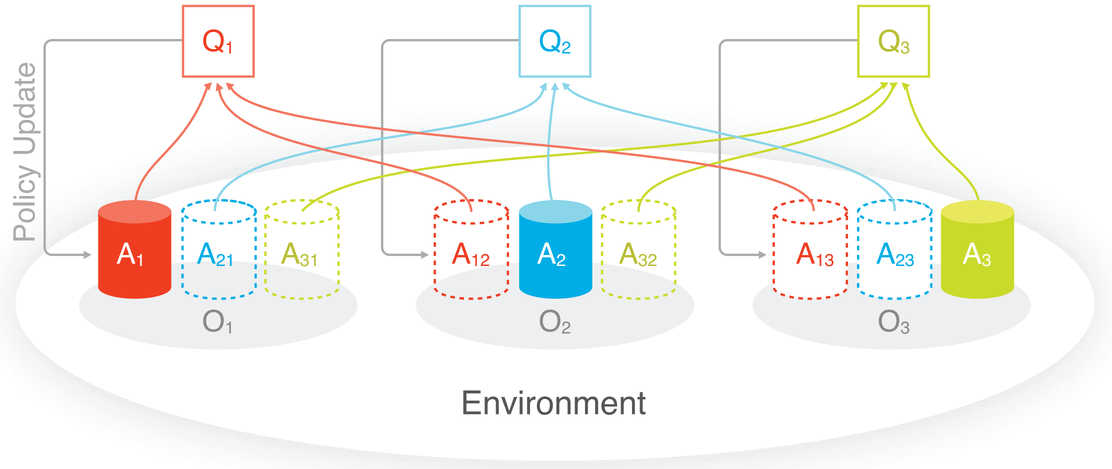

Figure 3.1: Illustration of MADDPG algorithm with three agents.

treating other participants as actual agents rather than inanimate objects within the environment. Each agent anticipates the actions of others in the field, a process that can be interpreted as a primitive version of Theory of Mind [Wel92]. Furthermore, the action evaluation of each agent is context-dependent, considering the states and predicted actions of all agents, which aligns with Tomasello's theory of coordinated hunting [TCC05]. Based on this assessment, agents subsequently adjust their policies to optimize their performance. Still, this planning remains individual-centric – while evaluating the collective context, agents are only concerned with their own actions. In this way, the algorithm serves as a competent representation for the MARL field and is the algorithm we adopted in the following case studies.

3.2 Case study 1: Free-rider effects

In the first case study, we explored how MARL behaves under the scenario with mixed interests [ZTD22]. Specifically, on top of the goal to catch the prey together, predators also incur individual action costs. Such setting thus poses a potential threat of the free rider problem to the coordination [Ols89]: Rational individuals benefit from the shared public goods even if they do not pay individual action costs. In the MADDPG approach towards the hunting task [LWT17], predators and prey have no action costs; thus, the free-rider problem is avoided altogether, since the only motivation for free-riding is to avoid individual costs in cooperation.

3.2.1 Experimental design

In the experiments, we systematically test MADDPG's performance in coordinated hunting with 4 experimental manipulations inspired by anthropological and animal studies.

3.2.1.1 Reward distribution among predators

Drawing inspiration from field observations of chimpanzee behavior where proximity to prey at the time of its demise plays a pivotal role in reward division [JDT19], we manipulate reward distribution among predators based on their distance-to-kill. Sensitivity to the distance-to-kill functions as a measure of selfishness: with a high index, rewards are concentrated around predators close to the kill, while a low index leads to broader dispersal. At extremes, pure selfishness results in the predator making the kill taking all the reward, while pure unselfishness leads to even distribution. All predators in a given condition adhere to the same mechanism. Formally, we define the reward distribution as an exponential function of the distance-to-kill, such that

$$R_i \propto (d_i + 1 - k)^{-s}, \tag{3.3}$$

where an agent i with d_i distance-to-kill receives R_i proportion of the reward, with selfish index s. The constant k denotes the minimum distance between two agents.

3.2.1.2 Action costs inducing the free-rider problem

Agents' inclination to free-ride during coordinated hunting is motivated by avoiding individual costs [Ols65, Ols89]. As action costs rise, agents prefer to remain static to lower these costs while still receiving allocated group rewards. To examine the free-rider problem's intensity in relation to reward distribution, we define the action costs to be proportional to the force exerted by agents, with the action cost for agent i, $C_i = a * F_i$, where F_i is the force exerted by agent i, and a denotes the action cost ratio in the specific condition. The action costs are applied to individual agents no matter which reward mechanism they take.

3.2.1.3 Group size

Numerous animal studies have shown that hunting party size is positively correlated with the success of the hunters across different species [MW99,SPM18,MTS14,Bed88,CC95]. We explore this correlation by testing various predator group sizes and their impact on hunting performance.

3.2.1.4 Hunting risks

Hunting risks influence animal behavior, with some species showing increased participation when risks are high [MTS14, New07]. We investigate this dynamic by altering the speed of prey, thereby manipulating hunting risks and studying their interplay with reward distribution.

3.2.2 Results

Our results demonstrated significant main effects for all four variables, with the selfish index showing significant interaction with the other three. An analysis of two-way interaction terms revealed that the performance of selfish agents (measured by kills per episode) increased with group size, while the performance of unselfish agents remained constant or declined with larger groups (Figure 3.2). Taking the group size of 6 as an example, without loss of generality, performance improved with increasing selfishness, with the most selfish predators achieving the highest performance. Action cost had a negative effect on performance, with a more pronounced effect on unselfish agents. The unselfish agents failed to gain any rewards when action cost reached .01, while selfish agents maintained high performance even under the largest action cost condition.

Our result strongly indicates the presence of the free-rider problem under the reward-sharing mechanism. The study thus highlights the modeling concern originated from the gap between MARL problem settings and the realistic settings in nature – agents working together not only care about the group benefits but also hold individual interests. In this case without an explicit enforcement of cooperation, agents modeled through the reward-based MARL abandoned the optimal Nash equilibrium of jointly hunting the prey together and settled for the local minima equilibrium to not move at all.

3.3 Case study 2: Joint commitment to one goal under distractions

In this case study, we explored whether cooperation modeled through MARL holds when there exists distractions [TGZ22]. The MADDPG algorithm was applied to a similar hunting environment with multiple predators but chasing multiple prey agents.

3.3.1 Experimental design

In the experiments, we investigate whether a group of three predators can demonstrate robust commitment when confronted with multiple prey options. The reward allocation mechanism remains constant, with an even distribution of rewards among predators upon capturing a prey. There are no associated action costs for any agent. We manipulate the quantity of prey

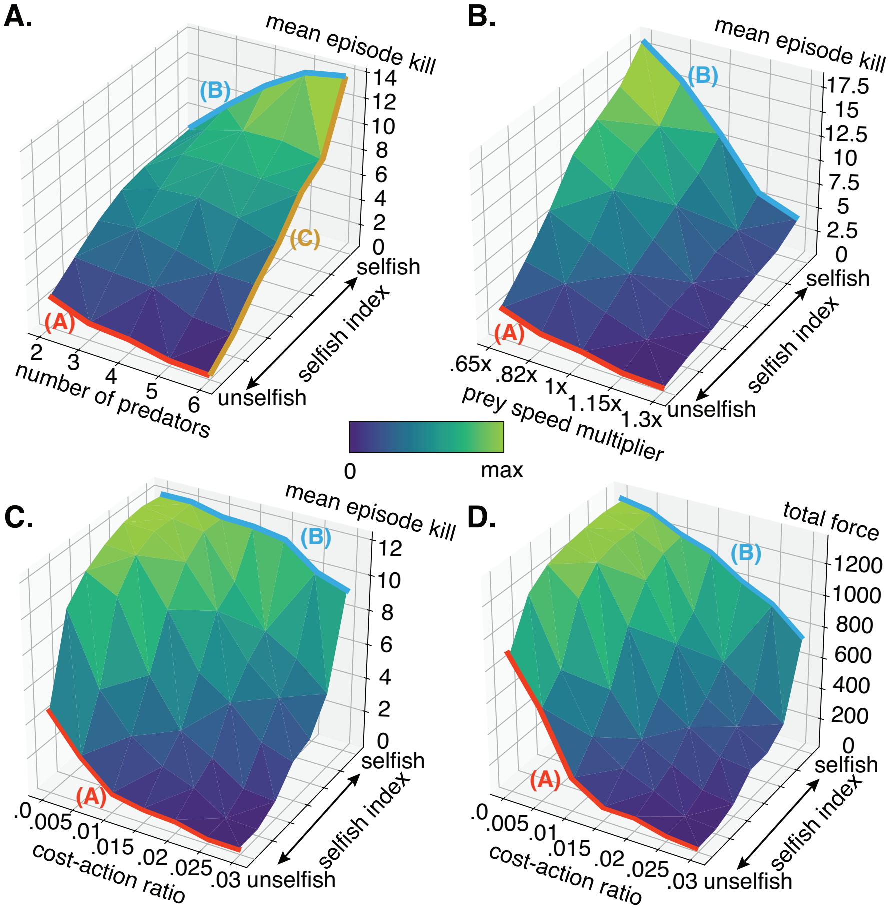

Figure 3.2: Performance of MADDPG in the coordinated hunting task.

to be 1, 2, 4. The game is structured such that the prey moves faster than the predators, necessitating a cooperative strategy where predators persistently chase a single prey. While there are no predefined targets, in order to improve performance and maximize cumulative rewards, the optimal strategy would be for the predators to hunt collectively, targeting one prey at a time.

3.3.2 Results

Our results indicate a significant main effect of prey group size on the task performance. Surprisingly, as the number of prey agents increases, the performance of predators gradually decreases – despite the increased opportunity to accumulate rewards due to a larger prey population. In this way, the simple incentive of sharing rewards does not suffice for the agents to coordinately learn to commit on one prey target to hunt. In order to maximize the reward function, the MADDPG agents turn to whichever prey that is closer to themselves without considering group intentions. Humans however can still coordinate efficiently when presented with a large number of potential targets [TGZ22].

CHAPTER 4

The Imagined We Framework

4.1 Motivation

Theories in cognitive science suggest that humans follow a shared agency perspective during cooperation, where members of a team form a “We” mind with joint mental states including a common-ground, joint attention, and joint intention to work on tasks as a whole [Tom09]. In this way, to form a shared intention, each agent individually demonstrates a readiness to commit to the joint goal through her behavior. During this process, collaborators view themselves not as individuals, but as part of an imagined, larger joint entity “We” acting under the shared beliefs, shared desires, and shared intentions of the group [Gil99]. With a “bird’s eye perspective”, a true central controller “We” with complete knowledge of the situation could perfectly coordinate all cooperators. However, in reality there is no central controller, and it is unrealistic and inefficient for all agents to share all knowledge. As a result, we call our framework “Imagined We” where each agent instead individually simulates a We agent, making We imagined. IW contains joint mental states given observation of joint actions already made by themselves and other agents [TSZ20, TGZ22].

4.2 Formulation

Our model builds upon the current progress of Bayesian modeling of shared intention. Under a Bayesian-theory-of-mind (BToM) framework named *Imagined We* (IW), collaborators view themselves not as individuals, but as part of an imagined, larger joint entity “We” acting under the shared beliefs, shared desires, and shared intentions of the group [Gil99] (Figure 4.1). Since there is no ground truth of "We" to infer, agents use the idea of bootstrapping with BToM inference to determine what "We" want to do by looking at what "We" have done:

$$\begin{aligned} &P(\text{“We” mind} \mid \text{Joint action, Environment}) \propto \\ &P(\text{Joint action} \mid \text{“We” mind, Environment})P(\text{“We” mind} \mid \text{Environment}) \end{aligned} \tag{4.1}$$

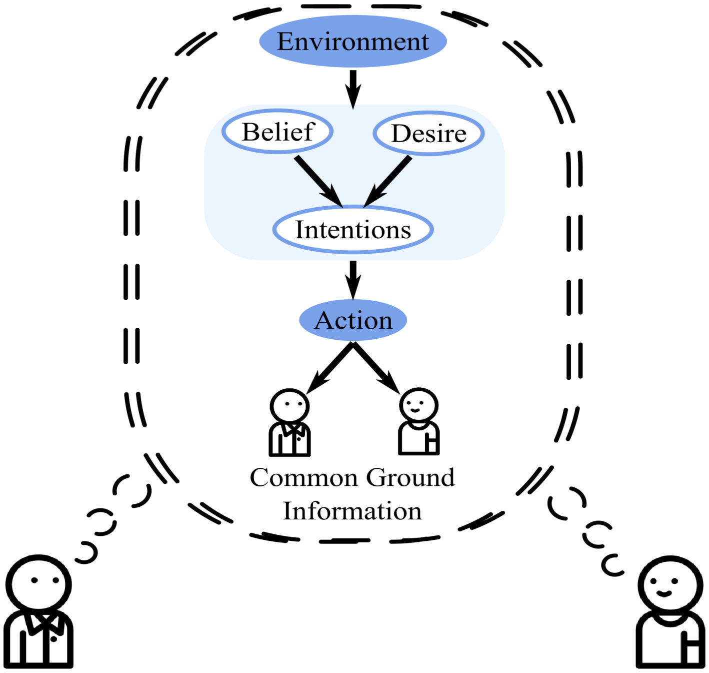

Figure 4.1: Imagined We representation.

In the context of the cooperative hunting task, the environment is fully observable without any uncertainty. The only uncertainty stems from the intention of "We" concerning "which prey should 'We' pursue persistently?" We simulate the inference of "We" intention through a three-step bootstrapping method: goal sampling, planning, and inference (Figure 4.2.) Specifically, the agents start by sampling one goal as the current goal of their version of the "We" agent – in this context, which prey they believe is the target based on their version of

IW. In the planning phase, given the goal, each agent acts by asking "what does 'we' expect me and others to do?" Aside from taking its own action following the intention of "We", an individual agent also expects others to take their actions as demanded by "We." In this way, each agent is simulating a centralized planner from its own perspective, with only one goal to pursue. In practice, this is achieved by applying the MADDPG algorithm which outputs a joint action, including the agent's own action to take as well as an expectation of other agents' actions. After taking one's own action based on the policy determined in the planning phase, each agent observes the actions actually taken by other agents. This enables a Bayesian ToM inference process: Conditioning on the observed actions, each predator computes the posterior probability of a given target being their joint goal. After updating the posterior of the Imagined We mind, each agent goes back to the first step, sampling a new goal and repeating the process. As more observations accumulate, agents further update their belief in the mind of "We" and gradually converge to a version of "We" that is treated as the mutual agreement among collaborators. The resulting "We" version thus serves as a social contract that agents aim to follow .

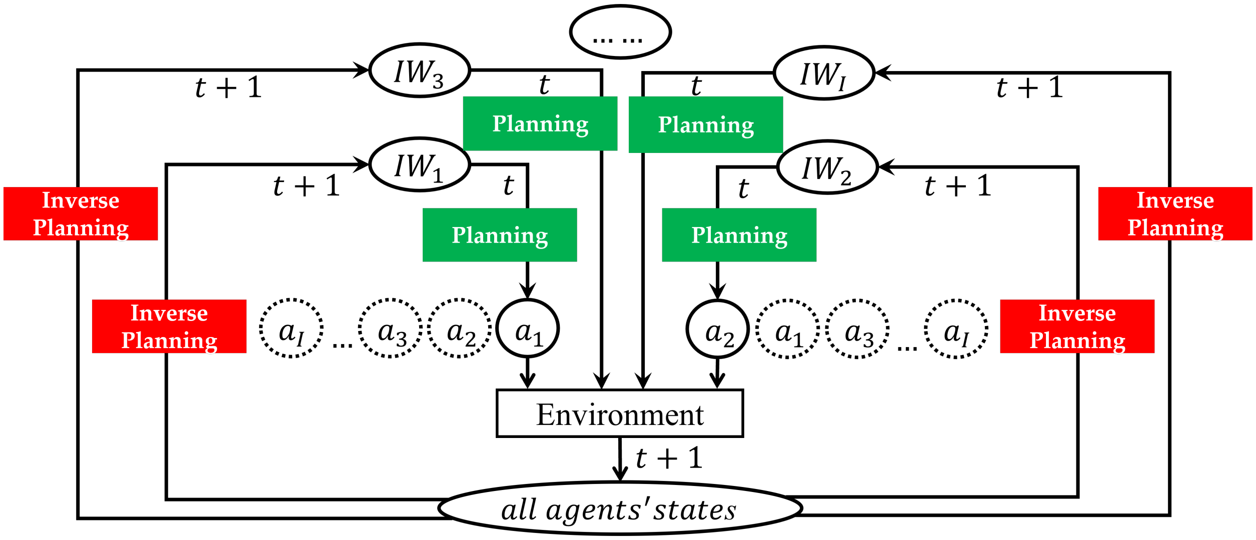

Figure 4.2: The framework of bootstrapping an Imagined We.

4.3 Experiments

We apply the IW model to a coordinated hunting scenario with three predators hunting 1, 2, or 4 prey agents, same setting as the study in section 3.3 [TGZ22]. Contrast to the result of MADDPG model (reward sharing, denoted by RS in the figure) that we found in the case study in section 3.3, the main effect of prey set size was not significant. The results revealed that the performance of the IW model did not decrease as the number of targets increased 4.3.

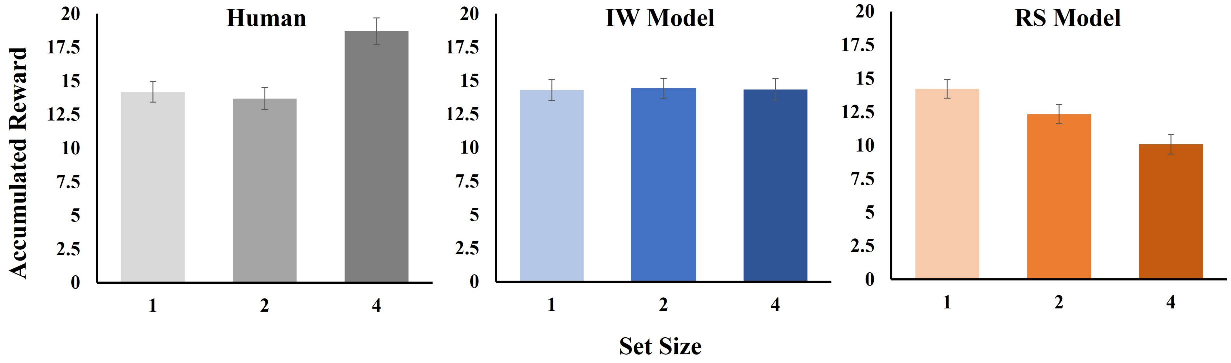

Figure 4.3: Model and human performance under different prey set sizes.

Besides the overall performance, we further assess hunting quality in humans, IW model, as well as a baseline MADDPG model by measuring the "duration of touch", which refers to the continuous time steps when at least one hunter is in contact with the prey. This duration is used as an indicator of the likelihood of a successful catch similar to real life scenarios. We classify the quality of raw rewards into three categories based on touch duration: low (1 time step), medium (2 time steps), and high (3 or more time steps) (Figure 4.4.) Results indicate that IW model demonstrated a relatively good quality of hunting and outperformed the MADDPG model in acquiring high-quality rewards, although it still has room for improvement as compared to human performance.

We further studied predators' goal consistency by measuring the entropy of the touched prey distribution, where lower entropy suggests a greater shared focus among the hunters

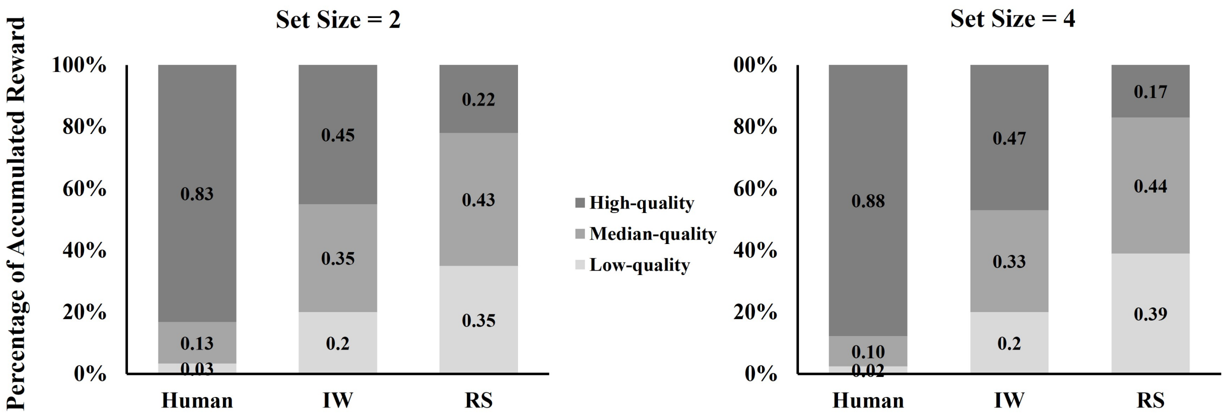

Figure 4.4: Results of the percentages of different quality rewards.

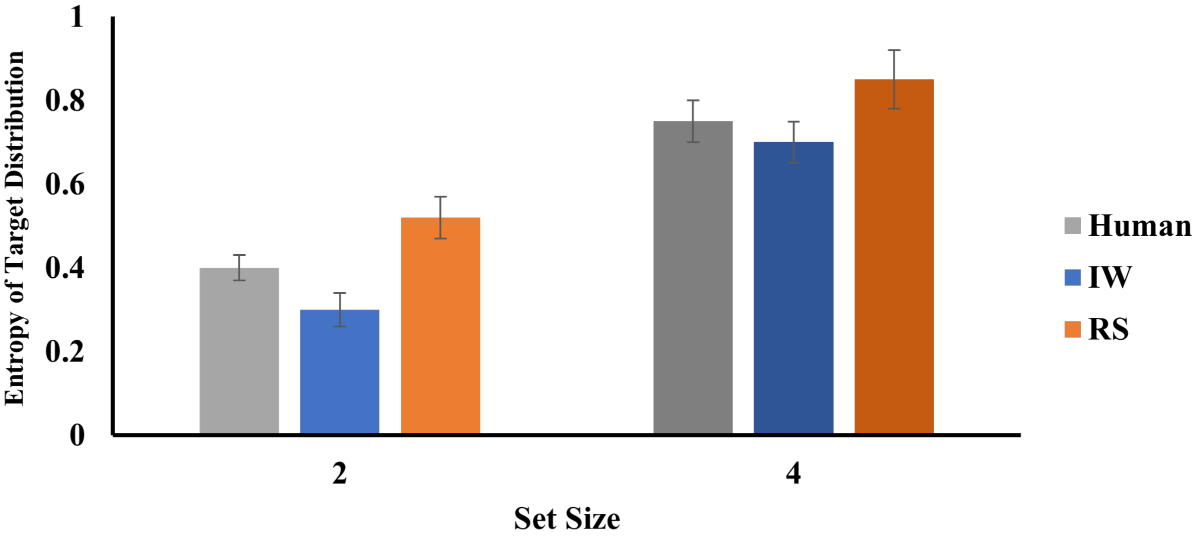

Figure 4.5: Results of the entropy of touched target distribution.

(Figure 4.5.) In both the set size 2 and 4 scenarios, the IW model demonstrated notable similarities with human goal entropy. Both the humans and the IW model consistently showed higher goal entropy than the MADDPG model, indicating that the IW model more effectively mirrors human behavior in goal pursuit than the MADDPG model.

In this way, cooperators modeled by IW were able to quickly bootstrap commitment to the same arbitrary intention without even the need for explicit communication. IW successfully captures humans' robust commitment in cooperation: resisting alternative targets, achieving greater quality of hunt, and maintaining a relatively high goal consistency among hunters. This indicates that there is indeed a bonus for agents to cooperate in tasks, and manifesting such bonus requires joint commitment as a stronger constraint for the team behavior.

4.4 Other applications

The IW framework has also been successfully applied to communication scenarios, where communication serves to coordinate perspectives, aligning "We" minds for better cooperation among agents [SLZ21]. Assuming agents to be rational and utility-maximizing cooperators, cooperation simulated by Imagined We framework successfully solves ambiguity in a signaling game (Figure 4.6.) In this cooperative task, a signaler and a receiver navigate a gridworld environment aiming to reach a target item, known only to the signaler. The signaler acts first, choosing to walk towards an item, send a signal indicating a single feature of the target (shape or color), or quit, while each step incurs a shared cost. The trial concludes either after the receiver's turn or when the target item is reached, with correct selection rewarding both agents, but traveling to the wrong item resulting in negative utility proportional to the steps taken. Under the framework of IW, the meaning of a signal is explicitly linked to the sender's intent, conveying information to resolve uncertainty about shared beliefs, desires, or actions. Results demonstrate that the IW model, compared to baseline models, shows superior performance under uncertainty without needing deep recursion.

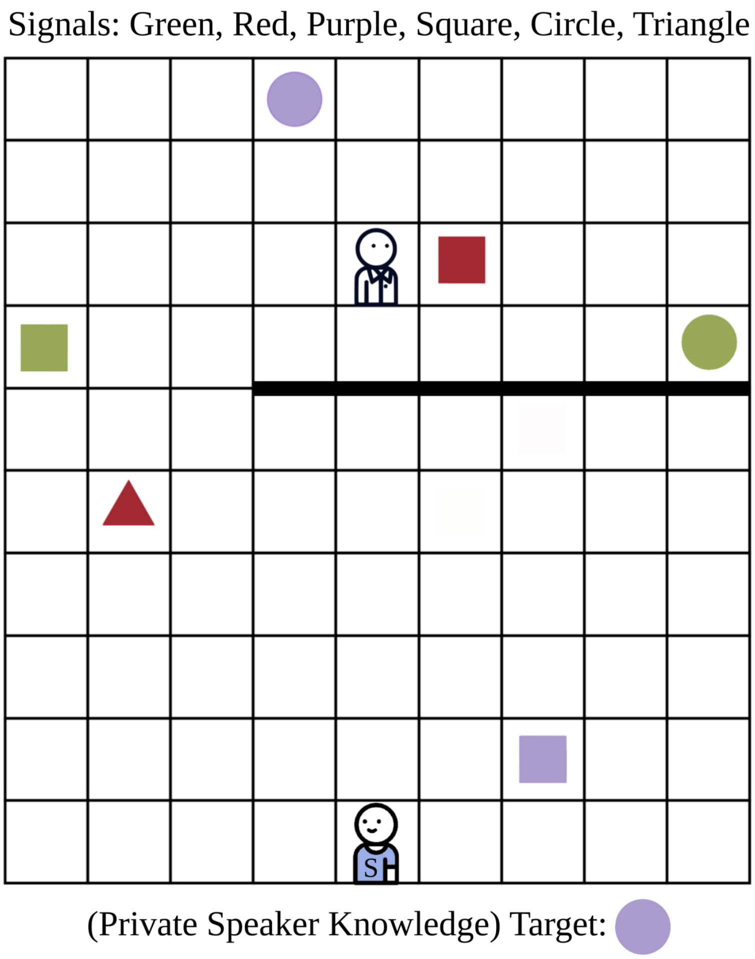
Signals: Green, Red, Purple, Square, Circle, Triangle
S
(Private Speaker Knowledge) Target: ●

Figure 4.6: Example trial setup in the signal communication game.

CHAPTER 5

Conclusions and Future Directions

Our work focuses on highlighting the differences between assumptions taken by the mainstream MARL models and cognitive theories in modeling cooperation. MARL enjoys the flexibility of being able to train agents with different social intentions through adjusting the relations between reward functions and defines cooperation as sharing the same objective. However, this assumption creates a disconnect with real-life human cooperation, where cooperative circumstances often encompass mixed interests.

Drawing on insights from cognitive science studies, we highlight canonical challenges identified by these theories that can potentially disrupt MARL training. In particular, within a coordinated hunting scenario, agents demonstrate an inability to learn effective cooperation when action costs come into play, signifying the presence of a free-rider effect. Moreover, when multiple prey serving as distractors are introduced into the environment, agents struggle to converge on a single prey for a joint hunt, leading to a decline in performance despite the increased availability of resources. In this way, to effectively mimic human-like cooperation, it is essential to consider modeling constraints that extend beyond the simple sharing of rewards. Inspired by the theory of shared intentionality in cognitive science, we introduce a novel modeling architecture *Imagined We*, where collaborators perceive themselves not merely as isolated individuals, but as components of a larger imagined entity, a "We", acting under the shared beliefs, desires, and intentions of the group. We show that this framework can enhance cooperation efficiency and more effectively mirror human behavior.

As we move forward, we will continue to refine and expand the *Imagined We* framework to facilitate more sophisticated cooperative behavior. When agents work together on a joint task, a host of critical factors come into play including fairness (determining how much resource each member is entitled to), ownership (identifying and respecting the property rights of others), and efficient role allocation (assigning roles that best align with individual skills and abilities). By modeling these essential concepts, we aim to deepen our understanding of the mechanisms underlying human cooperation. This in-depth knowledge will in turn equip us to design AI systems that can collaborate more efficiently and safely with humans in the future.

REFERENCES

[BBC19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. "Dota 2 with large scale deep reinforcement learning." *arXiv preprint arXiv:1912.06680*, 2019.

[BBD08] Lucian Busoniu, Robert Babuska, and Bart De Schutter. "A comprehensive survey of multiagent reinforcement learning." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **38**(2):156–172, 2008.

[Bed88] James C Bednarz. "Cooperative hunting Harris' hawks (Parabuteo unicinctus)." *Science*, **239**(4847):1525–1527, 1988.

[BR09] Robert Boyd and Peter J Richerson. "Culture and the evolution of human cooperation." *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1533):3281–3288, 2009.

[Bra92] Michael E Bratman. "Shared cooperative activity." *The philosophical review*, **101**(2):327–341, 1992.

[CC95] Scott Creel and Nancy Marusha Creel. "Communal hunting and pack size in African wild dogs, Lycaon pictus." *Animal Behaviour*, **50**(5):1325–1339, 1995.

[FFA18] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. "Counterfactual multi-agent policy gradients." In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[FNF17] Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. "Stabilising experience replay for deep multi-agent reinforcement learning." In *International conference on machine learning*, pp. 1146–1155, 2017.

[GBC09] Maria Gräfenhain, Tanya Behne, Malinda Carpenter, and Michael Tomasello. "Young children's understanding of joint commitments." *Developmental psychology*, **45**(5):1430, 2009.

[GCE05] Stefanie K Gazda, Richard C Connor, Robert K Edgar, and Frank Cox. "A division of labour with role specialization in group–hunting bottlenose dolphins (Tursiops truncatus) off Cedar Key, Florida." *Proceedings of the Royal Society B: Biological Sciences*, **272**(1559):135–140, 2005.

[Gil92] Margaret Gilbert. *On social facts.* Princeton University Press, 1992.

[Gil99] Margaret Gilbert. "Obligation and joint commitment." *Utilitas*, **11**(2):143–163, 1999.

[GNS09] Tao Gao, George E Newman, and Brian J Scholl. "The psychophysics of chasing: A case study in the perception of animacy." *Cognitive psychology*, **59**(2):154–179, 2009.

[Har68] Garrett Hardin. "The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality." *science*, **162**(3859):1243–1248, 1968.

[HBZ04] Eric A Hansen, Daniel S Bernstein, and Shlomo Zilberstein. "Dynamic programming for partially observable stochastic games." In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, volume 4, pp. 709–715, 2004.

[HS15] Matthew Hausknecht and Peter Stone. "Deep recurrent q-learning for partially observable mdps." In *2015 aaai fall symposium series*, 2015.

[HSB97] Kay E Holekamp, Laura Smale, R Berg, and Susan M Cooper. "Hunting rates and hunting success in the spotted hyena (Crocuta crocuta)." *Journal of Zoology*, **242**(1):1–15, 1997.

[HWG11] Katharina Hamann, Felix Warneken, Julia R Greenberg, and Michael Tomasello. "Collaboration encourages equal sharing in children but not in chimpanzees." *Nature*, **476**(7360):328–331, 2011.

[HWT12] Katharina Hamann, Felix Warneken, and Michael Tomasello. "Children's developing commitments to joint goals." *Child development*, **83**(1):137–145, 2012.

[HZA18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

[IS19] Shariq Iqbal and Fei Sha. "Actor-attention-critic for multi-agent reinforcement learning." In *International Conference on Machine Learning*, pp. 2961–2970. PMLR, 2019.

[JDT19] Maria John, Shona Duguid, Michael Tomasello, and Alicia P Melis. "How chimpanzees (Pan troglodytes) share the spoils with collaborators and bystanders." *PloS One*, **14**(9):e0222795, 2019.

[KST18] Ulrike Kachel, Margarita Svetlova, and Michael Tomasello. "Three-year-olds' reactions to a partner's failure to perform her role in a joint commitment." *Child Development*, **89**(5):1691–1703, 2018.

[Lee14] Peter T Leeson. *Anarchy unbound: Why self-governance works better than you think*. Cambridge University Press, 2014.

[LHP15] 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.” *arXiv preprint arXiv:1509.02971*, 2015.

[Lit94] Michael L Littman. “Markov games as a framework for multi-agent reinforcement learning.” In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

[LMB21] Yunzhe Liu, Marcelo G Mattar, Timothy EJ Behrens, Nathaniel D Daw, and Raymond J Dolan. “Experience replay is associated with efficient nonlocal learning.” *Science*, **372**(6544), 2021.

[LP02] Michail G Lagoudakis and Ronald Parr. “Learning in zero-sum team markov games using factored value functions.” *Advances in Neural Information Processing Systems*, **15**, 2002.

[LWT17] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. “Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, p. 6382–6393, Red Hook, NY, USA, 2017. Curran Associates Inc.

[LZL17] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. “Multi-agent reinforcement learning in sequential social dilemmas.” *arXiv preprint arXiv:1702.03037*, 2017.

[MFT15] Alicia P Melis, Anja Floedl, and Michael Tomasello. “Non-egalitarian allocations among preschool peers in a face-to-face bargaining task.” *PLoS One*, **10**(3):e0120494, 2015.

[MGK16] Alicia P Melis, Patricia Grocke, Josefine Kalbitz, and Michael Tomasello. “One for you, one for me: Humans’ unique turn-taking skills.” *Psychological science*, **27**(7):987–996, 2016.

[MKS15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. “Human-level control through deep reinforcement learning.” *Nature*, **518**(7540):529–533, 2015.

[MLL12] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. “Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems.” *The Knowledge Engineering Review*, **27**(1):1–31, 2012.

[MTS14] Daniel R. MacNulty, Aimee Tallian, Daniel R. Stahler, and Douglas W. Smith. “Influence of Group Size on the Success of Wolves Hunting Bison.” *PLOS ONE*, **9**(11):1–8, 11 2014.

[MW99] John C Mitani and David P Watts. "Demographic influences on the hunting behavior of chimpanzees." *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, **109**(4):439–454, 1999.

[New07] Nicholas E Newton-Fisher. *Chimpanzee hunting behavior*. Springer-Verlag, 2007.

[Ols65] Mancur Olson. *Logic of collective action: Public goods and the theory of groups (Harvard economic studies. v. 124)*. Harvard University Press, 1965.

[Ols89] Mancur Olson. "Collective action." In *The invisible hand*, pp. 61–69. Springer, 1989.

[OPA17] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. "Deep decentralized multi-task multi-agent reinforcement learning under partial observability." In *International Conference on Machine Learning*, pp. 2681–2690. PMLR, 2017.

[OSV08] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. "Optimal and approximate Q-value functions for decentralized POMDPs." *Journal of Artificial Intelligence Research*, **32**:289–353, 2008.

[PBA21] Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. "Using large-scale experiments and machine learning to discover theories of human decision-making." *Science*, **372**(6547):1209–1214, 2021.

[RSS18] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning." In *International Conference on Machine Learning*, pp. 4295–4304. PMLR, 2018.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[SC96] Tuomas W Sandholm and Robert H Crites. "Multiagent reinforcement learning in the iterated prisoner's dilemma." *Biosystems*, **37**(1-2):147–166, 1996.

[Sky04] Brian Skyrms. *The stag hunt and the evolution of social structure*. Cambridge University Press, 2004.

[SLA15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. "Trust region policy optimization." In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

[SLG17] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. "Value-decomposition networks for cooperative multi-agent learning." *arXiv preprint arXiv:1706.05296*, 2017.

[SLZ21] Stephanie Stacy, Chenfei Li, Minglu Zhao, Yiling Yun, Qingyi Zhao, Max Kleiman-Weiner, and Tao Gao. "Modeling communication to coordinate perspectives in cooperation." In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2021.

[SPM18] Liran Samuni, Anna Preis, Alexander Mielke, Tobias Deschner, Roman M Wittig, and Catherine Crockford. "Social bonds facilitate cooperative resource sharing in wild chimpanzees." *Proceedings of the Royal Society B: Biological Sciences*, **285**(1888):20181643, 2018.

[SSP21] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. "Reward is enough." *Artificial Intelligence*, **299**:103535, 2021.

[SSS16] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. "Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving." *CoRR*, **abs/1610.03295**, 2016.

[SSS17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. "Mastering the game of go without human knowledge." *Nature*, **550**(7676):354–359, 2017.

[SWD17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. "Proximal policy optimization algorithms." *arXiv preprint arXiv:1707.06347*, 2017.

[SYZ20] Wesley Suttle, Zhuoran Yang, Kaiqing Zhang, Zhaoran Wang, Tamer Başar, and Ji Liu. "A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning." *IFAC-PapersOnLine*, **53**(2):1549–1554, 2020.

[SZL21] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. "Decentralized Q-learning in zero-sum Markov games." *Advances in Neural Information Processing Systems*, **34**, 2021.

[Tan93] Ming Tan. "Multi-agent reinforcement learning: Independent vs. cooperative agents." In *Proceedings of the tenth international conference on machine learning*, pp. 330–337, 1993.

[TCC05] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. "Understanding and sharing intentions: The origins of cultural cognition." *Behavioral and Brain Sciences*, **28**(5):675–691, 2005.

[TGZ22] Ning Tang, Siyi Gong, Minglu Zhao, Chenya Gu, Jifan Zhou, Mowen Shen, and Tao Gao. "Exploring an Imagined "We" in Human Collective Hunting: Joint Commitment within Shared Intentionality." In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2022.

[TMK17] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. "Multiagent cooperation and competition with deep reinforcement learning." *PloS one*, **12**(4):e0172395, 2017.

[Tom09] Michael Tomasello. *Why we cooperate*. MIT press, 2009.

[Tom19] Michael Tomasello. *Becoming human*. Harvard University Press, 2019.

[TSZ20] Ning Tang, Stephanie Stacy, Minglu Zhao, Gabriel Marquez, and Tao Gao. "Bootstrapping an imagined we for cooperation." In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020.

[VBC19] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning." *Nature*, **575**(7782):350–354, 2019.

[VCT16] Amrisha Vaish, Malinda Carpenter, and Michael Tomasello. "The early emergence of guilt-motivated prosocial behavior." *Child Development*, **87**(6):1772–1782, 2016.

[WCT06] Felix Warneken, Frances Chen, and Michael Tomasello. "Cooperative activities in young children and chimpanzees." *Child development*, **77**(3):640–663, 2006.

[Wel92] Henry M Wellman. *The child's theory of mind*. The MIT Press, 1992.

[WLM11] Felix Warneken, Karoline Lohse, Alicia P Melis, and Michael Tomasello. "Young children share the spoils after collaboration." *Psychological science*, **22**(2):267–273, 2011.

[XCW20] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. "Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium." In *Conference on learning theory*, pp. 3674–3682. PMLR, 2020.

[YY10] Reuven Yosef and Nufar Yosef. "Cooperative hunting in brown-necked raven (Corvus rufficollis) on Egyptian mastigure (Uromastyx aegyptius)." *Journal of ethology*, **28**(2):385–388, 2010.

[ZKB20] Kaiqing Zhang, Sham Kakade, Tamer Basar, and Lin Yang. "Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity." *Advances in Neural Information Processing Systems*, **33**:1166–1178, 2020.

[ZTD22] Minglu Zhao, Ning Tang, Annya L Dahmani, Yixin Zhu, Federico Rossano, and Tao Gao. “Sharing Rewards Undermines Coordinated Hunting.” *Journal of Computational Biology*, 2022.

[ZYL18] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. “Fully decentralized multi-agent reinforcement learning with networked agents.” In *International Conference on Machine Learning*, pp. 5872–5881, 2018.