Uncertainty-Aware Opponent Modeling for Deep Reinforcement Learning

AAAI Track

Likun Yang
School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China
yanglikun2021@ia.ac.cn

Pei Xu
School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China
pei.xu@ia.ac.cn

Shiyue Cao
School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China
caoshiyue2021@ia.ac.cn

Yongjian Ren
School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China
renyongjian2022@ia.ac.cn

Xiaotang Chen
School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China
xtchen@nlpr.ia.ac.cn

Kaiqi Huang
School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
CAS Center for Excellence in Brain Science and Intelligence Technology
Beijing, China
xtchen@nlpr.ia.ac.cn

ABSTRACT

The ability to model opponent behavior is essential for autonomous decision-making in multi-agent games. Although stochastic behavior is universal in real-world situations, previous works have struggled to model opponents with high stochasticity, such as humans. The issue arises because stochasticity in opponent behavior introduces significant uncertainty into the opponent modeling process, which existing methods have not adequately addressed. We introduce a novel **U**ncertainty-**A**ware **O**pponent **M**odeling (**UAOM**) method that addresses two key sources of uncertainty stemming from the inherent randomness of the opponent's actions. The first pertains to the uncertainty in constructing the opponent model, while the second concerns the uncertainty in applying the model during decision-making. For the first uncertainty, UAOM uses a hybrid behavior modeling module to learn a more powerful opponent-aware representation by ensembling the deterministic and probabilistic models to address both aleatoric and epistemic uncertainties in opponent modeling. For the second uncertainty, UAOM uses an opponent-aware dynamic modeling module to learn a dynamic-aware representation. We further provide a theoretical analysis showing that jointly optimizing our two modules can enhance downstream reinforcement learning performance while ensuring system convergence. We evaluate UAOM in both simulated settings and human-agent interaction scenarios. Our experimental results show that the proposed method significantly enhances performance when facing opponents with varying degrees of stochastic behavior, while efficiently managing the uncertainties introduced by such opponents.

KEYWORDS

Opponent Modeling; Deep Reinforcement Learning; Uncertainty

ACM Reference Format:
Likun Yang, Pei Xu, Shiyue Cao, Yongjian Ren, Xiaotang Chen, and Kaiqi Huang. 2025. Uncertainty-Aware Opponent Modeling for Deep Reinforcement Learning: AAAI Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Opponent modeling refers to the task of understanding and predicting the intentions and behaviors of other agents (collectively termed as opponents, whether collaborators or competitors), which plays a vital role in autonomous decision-making systems [1]. A substantial body of research has focused on integrating opponent modeling with deep reinforcement learning, providing important insights into resolving issues such as partial observability [25], instability [37], and other challenges in reinforcement learning [22].

Existing methods [12, 25, 35] attempt to learn each step of an opponent's behavior as accurately as possible and integrate this precise model with reinforcement learning. However, building an accurate opponent model without errors is challenging, particularly when the opponent's behavior exhibits randomness. In real-world

scenarios, such randomness is prevalent [29], as humans often exhibit high levels of stochasticity due to operational errors or cognitive limitations, such as 'trembling hands' or 'fuzzy minds' [23]. The inherent randomness associated with such behaviors poses substantial challenges for the accurate modeling of opponents and consequently complicates reinforcement learning tasks that depend on opponent models.

In this paper, we introduce **U**ncertainty-**A**ware **O**pponent **M**odeling (**UAOM**), an effective method inspired by insights on uncertainty reduction, designed to address the challenges posed by opponents with highly stochastic behaviors. Such behaviors primarily introduce uncertainties in two critical processes: the modeling of opponents and the application of opponent models for reinforcement learning. UAOM tackles highly stochastic opponents by incorporating awareness of these two types of uncertainties.

The uncertainty during the modeling of opponents can be further categorized into two aspects: aleatoric uncertainty, stemming from the intrinsic randomness of opponent behavior, and epistemic uncertainty, arising from the cognitive difficulty of inferring opponent behavioral patterns. In contrast to existing works that typically employ a single deterministic network for opponent modeling [12, 25], lacking mechanisms to address uncertainty, UAOM employs a hybrid behavior modeling module that combines deterministic and probabilistic networks within an ensemble framework. The probabilistic network effectively handles the aleatoric uncertainty by capturing the inherent randomness in the opponent's actions, while the ensemble mechanism enhances the model's epistemic capacity to learn complicated opponent behaviors. The uncertainty during the application of the opponent model in downstream reinforcement learning tasks arises because the randomness of opponents' actions complicates the Markov dynamics faced by the controlled agent and intensifies the challenges of the reinforcement learning process. Existing opponent modeling approaches have largely overlooked the presence of this uncertainty [1, 35], thereby degrading the performance of downstream reinforcement learning tasks when faced with highly stochastic opponents. UAOM incorporates an opponent-aware dynamic modeling approach, which learns the environmental dynamics in the presence of opponents, effectively mitigating the uncertainty in employing the opponent model for reinforcement learning. We further provide theoretical analysis demonstrating that optimizing both proposed modules has the potential to enhance the performance of downstream reinforcement learning tasks while ensuring system convergence during the learning process.

To demonstrate the effectiveness of our method, we evaluate it in both simulated settings and human-agent interaction tasks [6]. We test the approach with simulated opponents showing varying levels of randomness [19] in three types of games: pure cooperation, pure competition, and mixed motives [7]. Results consistently show that our method outperforms the baseline, especially under high randomness conditions. Further validation through human-agent interaction tasks confirms its effectiveness in scenarios involving human opponents. Additionally, we conduct visualization experiments and ablation studies to demonstrate the effectiveness of managing uncertainty.

In summary, we contribute the following: (i) We introduce UAOM, which uses hybrid behavior modeling and opponent-aware dynamic modeling to address uncertainties in both opponent model construction and utilization. (ii) We present a theoretical analysis that illustrates the joint optimization of both proposed modules can improve the performance of downstream reinforcement learning tasks, while ensuring the convergence of the system. (iii) We show empirically that UAOM outperforms baselines in handling opponents with varying stochasticity in both simulated environments and human-agent interactions. We further analyze the effectiveness of UAOM in handling uncertainty.

2 RELATED WORK

2.0.1 Modeling Fixed-strategy Opponents. While deep reinforcement learning has yielded many effective methods[33, 34], opponent modeling nevertheless remains one of the most crucial ones[1, 36]. Many works have applied opponent modeling to cope with fixed-strategy agents. For example, He et al. [11] made a pioneering contribution by introducing agent modeling within deep reinforcement learning, reconstructing the opponent's actions using Deep Q-networks [21], which enabled the agent to predict actions during reinforcement learning. Raileanu et al. [27] inferred the opponent's intention by assuming the consistency of the opponent's policy and the controlled agent's policy. Hernandez-Leal et al. [12] leveraged agent modeling as an auxiliary task in reinforcement learning, and improved the performance when facing fixed-strategy opponents. Recurrent VAEs [24, 38] encoded a compact variational embedding of previous interactions with the agent, which conditioned the main agent policy. Papoudakis et al. [25] proposed an autoencoder architecture to model rational agents in partially observable environments. Yuan et al. [36] proposed employing in-context learning to model opponents within the offline reinforcement learning framework. These studies assume that the opponent follows a fixed strategy and do not address the potential stochasticity in the opponent's behavior. Compared to these works, our research primarily addresses the challenge of modeling opponents whose behaviors exhibit significant stochasticity.

2.0.2 Modeling Adaptive Opponents. Agent modeling in repeated games has attracted some researchers, who have focused on the update of agent policies. Rabinowitz et al. [26] proposed TomNet, a meta-learning approach to study agents with fixed and learnable policies in trajectory prediction environments. Lu et al. [18] developed a model-free method that used meta-learning to tackle the problem of opponent policy updates. Foerster et al. [9] derived a model of opponent parameter update in reinforcement learning and improved the cooperation with opponents who updated their strategies. Khan et al. [15] extended the above model to more complex games and demonstrated its effectiveness. These studies primarily focus on agents with updating models, rather than addressing the inherent stochasticity in opponents' strategies. Some researchers have adopted hierarchical cognitive models [5] to model agents with reasoning ability. PR2 [32] was the first to introduce probabilistic recursive reasoning in DRL with a variational Bayes method. GR2 [31] modeled bounded rational agents using a recursive reasoning framework based on probabilistic graphical models. MBOM [35] developed an agent modeling algorithm that performed recursive

reasoning in the environment model and was able to cope with reasoning opponents. These studies model bounded rationality using recursive reasoning models but do not account for bounded rational behavior patterns with inherent randomness, such as quantal response [19]. In contrast to these works, which focus on opponent behavior models with self-updating capabilities or reasoning abilities, our research investigates opponents with a different behavioral pattern, akin to the quantal response model. Specifically, we focus on the high stochasticity that may exist in the opponent's behavior pattern, which is more commonly encountered in real-world applications.

3 PRELIMINARIES

Formally, a Partially Observable Stochastic Game (POSG) [10, 28] is a tuple $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, \mathcal{R} \rangle$, where $\mathcal{I} = \{1, 2, \ldots, n\}$ is the set of n agents; $\mathcal{S}$ is the set of states; $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$ is the set of joint actions; $\mathcal{O}$ is the set of states; $\mathcal{O} = \mathcal{O}_1 \times \mathcal{O}_2 \times \cdots \times \mathcal{O}_n$ is the set of joint observations; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function that determines how the state evolves. For each agent $i \in \mathcal{I}$ a reward function is $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, and access to its observation $o_i \in \mathcal{O}_i$ the observation function $\Omega_i : \mathcal{S} \times \mathcal{A} \times \mathcal{O}_i \to [0, 1]$ defines a probability distribution over the possible next observations of agent i given the previous state and the joint action of all agents.

The goal of the agent i we controlled is to maximize its expected cumulative discount rewards interacting with other agents $-i$

$$\mathbb{E}_{(s_i^t, a_i^t, a^t a_{-i}) \sim \mathcal{T}, \pi_i, \pi_{-i}} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s^t, a_i^t, a_{-i}^t) \right] \tag{1}$$

where $\gamma \in (0, 1)$ is a discount factor.

4 METHOD

We propose a novel approach called **U**ncertainty-**A**ware **O**pponent **M**odeling (**UAOM**), which addresses uncertainty both in the opponent modeling process and in the application of the opponent model. The Hybrid Behavior Modeling Module employs an ensemble of deterministic and probabilistic networks to address aleatoric and epistemic uncertainties during the modeling of opponents. This ensemble approach effectively captures the opponent's intrinsic randomness through the probabilistic network, while the ensemble of the deterministic network enhances the model's ability to infer diverse and complex behavioral patterns. Moreover, UAOM employs opponent-aware dynamic modeling to capture the transition dynamics influenced by the opponent's actions, providing consistent features that enhance the reinforcement learning process and mitigate uncertainty when utilizing the opponent model. The proposed approach also includes a theoretical foundation to ensure that the joint optimization of these components not only enhances policy performance but also guarantees convergence in reinforcement learning.

4.1 Hybrid Behaviour Modeling

To effectively address the uncertainty inherent in opponent modeling, we propose the Hybrid Behaviour Modeling Module as a component of the UAOM framework. This module is designed with awareness of two types of uncertainty: aleatoric uncertainty, caused by the inherent randomness in opponent behavior, and epistemic uncertainty, arising from the cognitive challenge of understanding the opponent's highly stochastic behavior. Unlike previous works that primarily focus on accurately predicting opponent actions, we draw inspiration from prior research [4, 17] on addressing randomness in modeling objectives. Rather than predicting a single deterministic output, we employ a network that models the opponent's behavior as a distribution. While this probabilistic network effectively mitigates aleatoric uncertainty, it does not inherently improve, and may even degrade, the model's cognitive capacity for capturing complex behaviors. To overcome this limitation, we introduce a novel ensemble framework that combines deterministic and probabilistic networks. This ensemble approach enhances the model's ability to cognitively represent opponent strategies, thereby alleviating the effects of epistemic uncertainty.

We assume the opponent's behavioral embedding z_t^B leverages the history trajectory of controlled agent $h_i^t = (a_i^{0:t-1}, o_i^{1:t})$ and the history trajectory of the modeled agent $h_{-i}^t = (a_{-i}^{0:t-1}, o_{-i}^{1:t})$ to infer the policy of modelled agent. z_t^B can be obtained by simultaneously learning encoder $z_t^B = f_{\phi^e}(h_i^t)$ and decoder $h_{-i}^t = f_{\phi^d}(z_t^B)$.

Specifically, the model is composed of a deterministic network and a probabilistic network. To enable adaptive representation fusion, they share the same recurrent encoder f_{ϕ^e} to learn z_t^B. The decoder consists of both a deterministic decoder and a probabilistic decoder, where $f_{\phi^d} = \{f^{de}_{\phi^d_{de}}, f^{pr}_{\phi^d_{pr}}\}$. The deterministic decoder predicts the exact trajectory, while the probabilistic decoder predicts a formalized distribution of the trajectory. In particular, with respect to the probabilistic decoder, we employ a Gaussian distribution $\mathcal{N}(z^Q; \mu_o, \sigma)$ to model the observation and a Gumbel distribution $\mathcal{F}(z^Q; \mu_a, I)$ to model the action. Furthermore, we obtain a reparameterized sampling [13, 16] process to sample observation and action respectively from the Gaussian distribution and Gumbel distribution, which makes the process trainable.

Both the observation output of the deterministic network and probabilistic network can be estimated by mean square error.

$$\mathcal{L}_{B_o} = \frac{1}{T} \sum_{t=1}^{T} (f^o_{\phi^d}(z_t^B) - o_{-i}^t)^2 \tag{2}$$

Both the action output of the deterministic decoder and probabilistic decoder can be estimated by negative log error. This form is particularly suitable because the action is represented as a distribution rather than a single-point estimate. By minimizing the negative log-likelihood, the model aligns its predicted distribution with the true action distribution.

$$\mathcal{L}_{B_a} = \frac{1}{T} \sum_{t=1}^{T} (-\log f^a_{\phi^d}(a_{-i}^t | z_t^B)) \tag{3}$$

4.2 Opponent-aware Dynamic Modeling

In downstream reinforcement learning, the agent operates within a partially observable Markov decision process. The behavior of opponents exerts substantial influence on the Markov dynamics encountered by the controlled agent, thereby complicating the policy learning process. This complexity introduces uncertainty into the process of leveraging the opponent model for more informed

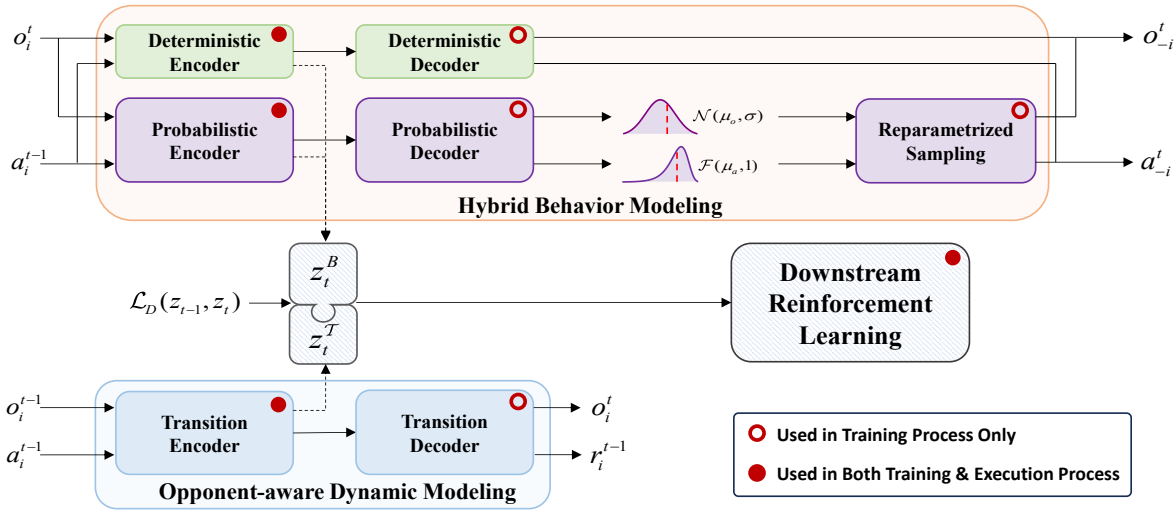

Figure 1: The Framework of UAOM. Unlike existing methods that rely solely on deterministic encoders, our innovative *hybrid behavior modeling* module combines deterministic and probabilistic networks to jointly model opponent behavior. Additionally, our original *opponent-aware dynamic modeling* module captures the transition dynamics influenced by the opponent.

decision-making. Specifically, this uncertainty is the subjective uncertainty about the dynamics function, due to a lack of sufficient information to uniquely determine the underlying system exactly. Unlike existing work that has not addressed this issue, we propose an opponent-aware dynamic modeling module to capture the information from dynamic transitions, thereby alleviating the impact of uncertainty in downstream reinforcement learning.

In partially observable environments, we model the subsequent feedback from the environment arises from the controlled agent's action, which implicitly provides information about the transition dynamics affected by opponents. Therefore, we define the opponent-aware dynamic function as $\mathcal{F}_{o'} = \{f : \mathcal{O} \times \mathcal{A} \times \mathcal{O} \to \mathcal{R}\}$. We use an encoder-decoder structure, where encoder is $z_t^{\mathcal{T}} = f_{\varphi^e}(\{o_i^{t-1}, a_i^{t-1}\})$ and decoder is $\{o_i^t, r_i^{t-1}\} = f_{\varphi^d}(z_t^{\mathcal{T}})$. Specifically, we adopt a recurrent encoder to maintain historical information, and a linear neural network to decode. Both state and reward can be estimated by mean square error.

$$\mathcal{L}_{\mathcal{T}} = \frac{1}{T}\sum_{t=1}^{T}\left[(f^o_{\varphi^d}(z_t^{\mathcal{T}}) - o_i^t)^2 + (f^r_{\varphi^d}(z_t^{\mathcal{T}}) - r_i^{t-1})^2\right] \tag{4}$$

4.3 Overall Training and Execution

During training, we assume access to the global historical information and update the behavioral model and the transition model accordingly. During execution, we only leverage the available partial observation, self-action history, and the learned encoder to infer the current belief z_t, where $z_t = z_t^B \oplus z_t^{\mathcal{T}}$ ($\oplus$ means concatenation in latent space). To ensure the stability of the latent embedding transferred to the downstream task avoiding the instability of the downstream reinforcement learning training, we devise a regularization loss to stabilize the belief optimizing process

$$\mathcal{L}_D = \beta\frac{1}{T}\sum_{t=1}^{T}[D_{KL}(p(z_{t-1})||p(z_t))] \tag{5}$$

where $p(z_t)$ is the distribution of embedding at step t, β is a hyperparameter to restrict the regularization loss, $D_{KL}(\cdot||\cdot)$ is the KL divergence operator. In our experiments, we adopted A2C [20] as the downstream reinforcement learning method, and defined the loss from reinforcement learning as $\mathcal{L}_r(\theta_r)$. Then, all the parameters $\theta = (\phi_e, \phi_d, \varphi_d, \varphi_e, \theta_r)$ can be optimized by the overall optimization objective:

$$\mathcal{L}(\theta) = \mathcal{L}_B + \mathcal{L}_{\mathcal{T}} + \mathcal{L}_D + \mathcal{L}_r \tag{6}$$

4.4 Theoretical Analysis

The proposed UAOM framework introduces a novel integration of the opponent behavior model with the dynamic model, jointly optimized alongside the reinforcement learning process. In this section, we undertake a theoretical analysis to explore the properties of the overall system resulting from this integration. Specifically, our goal is to address two fundamental questions: (1) Does the integration of the opponent behavior model and the dynamic model have the potential to enhance the performance of downstream reinforcement learning tasks? (2) Does coupling the proposed framework with reinforcement learning theoretically ensure convergence to an optimal solution?

In POSG mentioned in Section 3, we define the belief regions of the controlled agent as a finite set $\mathcal{Z}_i$, where $\mathcal{Z}_i : \mathcal{O}_i \times \mathcal{A}_i \times \mathcal{S} \to [0, 1]$. Thus, any belief $z_i \in \mathcal{Z}_i$ is defined as the probability distribution of the state s given the history of actions and observations of the controlled agent. With belief defined, we can formulate the decision

problem of controlled agent i to a tuple $< \mathcal{A}_i, \mathcal{O}_i, \mathcal{Z}_i, \mathcal{W}_i, \mu_i, \pi_i >$, where $\mathcal{W}_i$ are belief transition function with $\mathcal{W}_i(z_i, a_i, o_i', z_i')$ denoting the probability of transiting from z_i to z_i' when taking action a_i in belief z_i results in observing o_i', μ_i is the initial distribution of decision states with $\mu_i(z_i)$ denoting the probability of initially being in belief z_i, π_i are state-dependent stochastic policies with $\pi_i(z_i, a_i)$ denoting the probability of taking action a_i in belief z_i. Let $\theta_i = \{\pi_i, \mu_i, \mathcal{W}_i\}$ denote the parameters of the controlled agent i. Specially in the setting we studied, $\mathcal{W}_i$ is related to the transition $\mathcal{T}$ and policy of other agent $\pi_{-i}(a_{-i}|o_{-i})$. Since the decision is based on agent i, for simplicity, we omit the subscript i in the following expressions of this subsection. Under our setting, it is easy to observe a fact that the fact shows the feasibility that a Markov decision problem can be represented by the historical information and belief-based policy characterized by the parameter θ in our settings. This fact is formalized and proven in Appendix A.

Theorem 4.1. *Let $\mathcal{D}^{(K)}$ be a set of episodes obtained K trajectories by controlled agent i interacting with the environment by arbitrary stochastic soft policy π parameterized by θ, the expected sum of discounted rewards equals to $\lim_{K\to\infty} V(\mathcal{D}^{(K)}; \theta)$, where*

$$V(\mathcal{D}^{(K)}; \theta) = \frac{1}{K} \sum_{k=1}^{K} \sum_{t=0}^{T_k} \frac{\gamma^t r_t^k}{\prod_{\tau=0}^{t} p^{\pi}(a_{\tau^k}|h_\tau^k)} \prod_{\tau=0}^{t} p(a_\tau^k|h_\tau^k, \theta) \quad (7)$$

Theorem 4.1 is proven in Appendix B. Theorem 4.1 confirms that the expected sum of discounted rewards can be parameterized and optimized by θ, where $\theta = \{\pi, \mu, \mathcal{W}\}$. Revisiting the formulation, $\mathcal{W}$ encapsulates both the dynamic model and the opponent behavior model, demonstrating that both models directly influence the reward maximization process. This underscores that integrating the behavior model with the dynamic model improves the agent's policy during reinforcement learning. Hence, Theorem 4.1 substantiates the potential for performance improvement in downstream reinforcement learning tasks, addressing Question (1) affirmatively.

Theorem 4.2. *Let $\{\theta^{(1)}\theta^{(2)} \dots \theta^{(n)}\}$ be a sequence yielded by iteratively applying rules formalized in Appendix B. Then*

$$\lim_{K\to\infty} V(\mathcal{D}^K, \theta) \quad (8)$$

exists and the limit is a maxima of $V(\mathcal{D}^K, \theta)$.

The formal statement and proof of Theorem 4.2 are provided in Appendix C. Theorem 4.2 demonstrates that the iterative optimization process defined by the parameter θ converges to the maximum of the expected cumulative discounted reward V^* as the number of episodes $K \to \infty$. This addresses Question (2), confirming that our framework ensures the convergence to an optimal solution of the reinforcement learning process.

These two questions together establish the sufficiency of our framework. By integrating the behavior model and the dynamic model while systematically reducing their uncertainties, our approach ensures both improved performance and convergence in downstream reinforcement learning tasks. Specifically, Theorem 4.1 demonstrates that jointly optimizing these models leads to enhancements in policy performance, while Theorem 4.2 provides a guarantee that the reinforcement learning process converges as the accuracy of these models improves. Central to our framework is the systematic reduction of uncertainties in the behavior and dynamic models, which directly enhances their predictive accuracy and, by extension, the performance of the reinforcement learning system. This joint optimization strategy represents a departure from conventional approaches that focus solely on the behavior model [11, 12, 25], offering a comprehensive solution that explicitly addresses the interplay between these two models. Therefore, our theoretical analysis directly demonstrates that the UAOM framework is sufficient to improve performance and ensure convergence in reinforcement learning tasks. By demonstrating the critical role of uncertainty reduction in both models, our analysis validates the framework's design and highlights its effectiveness in achieving stable learning outcomes.

5 EXPERIMENTS

We conducted experiments in both simulated settings and with human-agent interactions. We assessed the method using opponents with varying levels of randomness across different game types, including pure cooperation, pure competition, and mixed motives. Our findings indicate that the method consistently outperforms the baseline, particularly in high-randomness scenarios. Validation with human opponents further supported its effectiveness. Additionally, we performed visualization and ablation studies to illustrate the effectiveness of managing the uncertainty.

5.1 Evaluation with Simulated Agents

5.1.1 Environmental Setup. We evaluated our method on three types of partially observable games: cooperative (double speaker-listener), mixed-motive (level-based foraging), and competitive (predator-prey). Double speaker-listener requires color-based cooperation, while level-based foraging involves food collection with competitive/cooperative elements. Predator-prey is a pursuit-evasion competing game. We categorized tasks for each game type based on different settings (e.g., hard/easy for double speaker-listener, large/scarcity for foraging, complicated/simple for predator-prey). Detailed settings are in Appendix D.

5.1.2 Opponent Setup. To simulate human opponents with different degrees of stochastic behavior, we use the widely recognized human behavior model, Quantal Response. This model features a hyperparameter λ, which can be adjusted continuously between 0 and ∞. When $\lambda = 0$, the opponent's behavior is entirely random, whereas when $\lambda = \infty$, the behavior becomes deterministic. To be more specific, when $\lambda = 1$, opponents have an average probability of 60.4% to "tremble" and select an unexpected action. At $\lambda = 2$, this probability drops to 35.1%, and by $\lambda = 7$, it further decreases to 0.43%. In each scenario, we have configured 10 distinct types of opponents that follow this behavioral model, and each opponent's hyperparameter λ can be adjusted to demonstrate varying levels of stochasticity. More detailed opponents' settings are provided in Appendix E.

5.1.3 Baselines. We compare our method against four baselines: **N**o **A**gent **M**odelling (**NAM**) is a reinforcement learning algorithm that does not employ any agent modeling techniques and relies on local observations and historical latent information, resembling RL2 [8]. **L**ocal **I**nformation **A**gent **M**odelling (**LIAM**) [25] adopts a

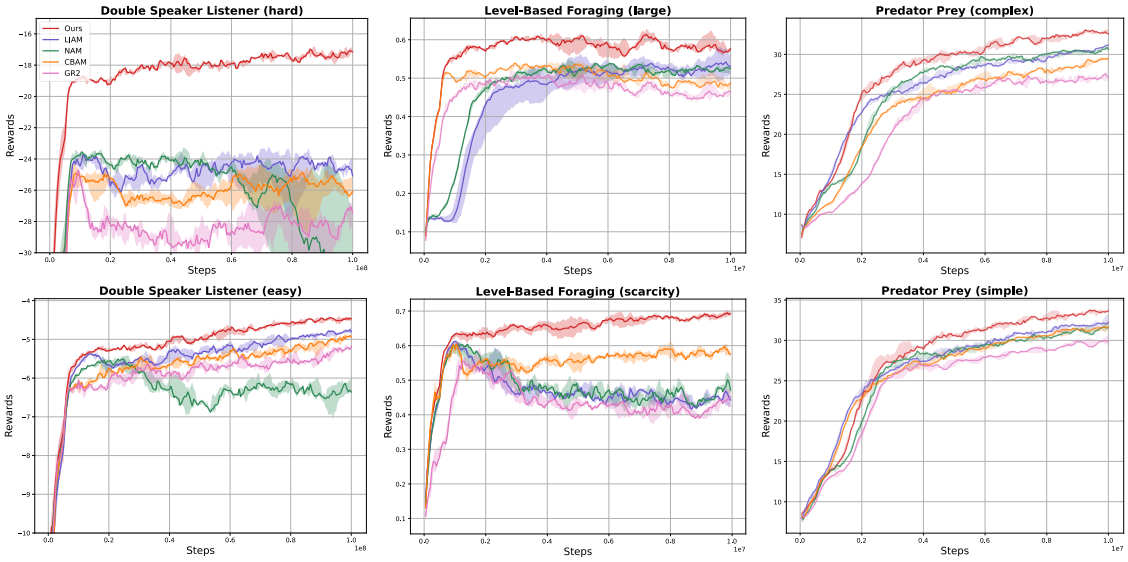

Figure 2: Performance with High-Stochasticity Opponents. Episodic evaluation rewards and 95% confidence intervals for the five evaluated methods. The results demonstrate that UAOM outperforms other baselines when dealing with highly stochastic opponents, as represented by the setting $\lambda = 1$.

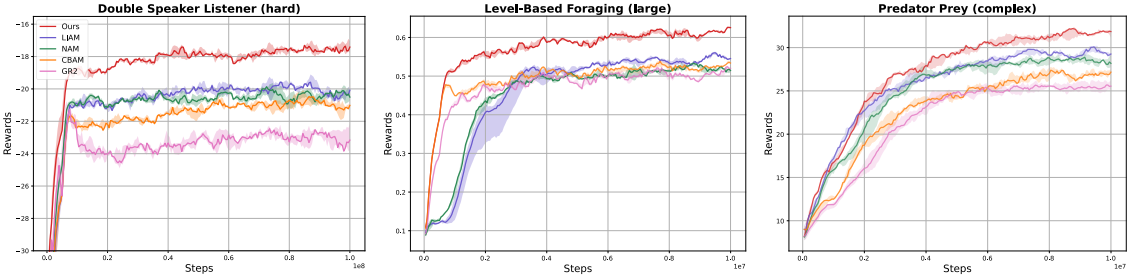

Figure 3: Performance with Lower-Stochasticity Opponents. Episodic evaluation rewards and 95% confidence intervals for the five evaluated methods. The results demonstrate that UAOM outperforms other baselines when dealing with lower stochastic opponents, as represented by the setting $\lambda = 2$.

representation learning approach to characterize the trajectories of agents with fixed strategies in partially observable environments without considering bounded rationality. It is a state-of-the-art method in agent modeling and has consistently served as a baseline for opponent modeling. **C**lassification-**B**ased **A**gent **M**odelling (**CBAM**) is derived from the context learning process of the algorithm [14]. It is an agent modeling algorithm that classifies policy identity by using the observations and actions of the modeled agent as inputs. During the training process, the algorithm optimizes by maximizing the log-likelihood of the policy identity, ultimately yielding the corresponding policy identity as output. **G**eneralized **R**ecursive **R**easoning (**GR2**) [31] is a reinforcement learning algorithm for modeling bounded rational agents. It assumes that agents possess varying degrees of reasoning rationality and utilizes k-order recursive reasoning to model the hierarchy of agents' rationality. This baseline enables higher-level agents to more effectively respond to agents with different levels of rationality.

5.1.4 Results in Opponents with High Stochasticity. We conducted an evaluation of our method by comparing it with baseline approaches at a hyperparameter setting of $\lambda = 1$. The training curves are included in Figure 2. Our method consistently surpassed all baseline methods across various scenarios. In the Double Speaker Listener (Hard) task, the randomness in the opponents' information output posed significant challenges for baseline methods, which failed to develop effective cooperative strategies. Notably, the NAM method struggled with convergence. In contrast, our approach effectively managed the behavioral uncertainty introduced by opponents, demonstrating not only superior rewards but also enhanced efficiency and reduced performance variance. Similarly, in the Level Based Foraging (Scarcity) task, baseline methods encountered difficulties in distinguishing between cooperative and competitive opponent behavior, often converging to suboptimal strategies. Our method successfully addressed this subjective uncertainty, leading to stable and substantial improvements in reward acquisition.

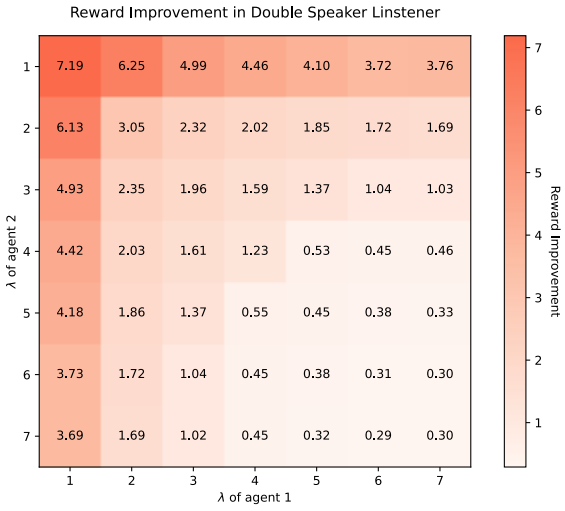

Figure 4: Performance Improvement Heatmap Across Varying Stochasticities. This heatmap shows the performance improvement of our method compared to LIAM in terms of reward when facing two opponents with varying levels of stochasticity. Our method effectively handles opponents with mixed stochasticity, demonstrating increasing effectiveness as the stochasticity rises.

5.1.5 Results in Opponents with Varying Stochasticity. To assess the adaptability and effectiveness of our approach, we evaluated its performance against opponents exhibiting different levels of stochasticity. The training curves in Figure 3 illustrate the results for scenarios with a hyperparameter setting of $\lambda = 2$, where our method consistently outperforms baseline approaches across all tasks. By comparing the outcomes in Figures 2 and 3, we observe that as opponent stochasticity increases, the performance gap between our method and the baseline widens. To delve deeper into how our method adapts to varying levels of opponent randomness, we created a speaker-listener environment with two opponents, each characterized by different levels of stochasticity. Figure 4 presents the results, showing that our method is capable of managing multiple opponents with diverse stochastic behaviors. Notably, our method delivers performance improvements in both low and high stochasticity settings, with more significant gains observed as the level of opponent randomness increases.

5.2 Evaluation in Human-agent Interaction

5.2.1 Experimental Setup. Overcooked is a cooperative game where two players, each controlling a chef, work together to cook and deliver dishes within a time limit. It is a widely recognized benchmark for human-agent collaboration [6], used to assess agents' interaction with human players. We evaluated three challenging sub-environments—Cramped Room, Asymmetric Advantages, and Coordination Ring—using 10 distinct human proxies generated from real gameplay data for each. Details are in Appendix F. We used the same baselines as in Section 5.1, but the GR2 method failed to converge in the high-dimensional Overcooked environment and was excluded from comparisons. Our method also outperforms

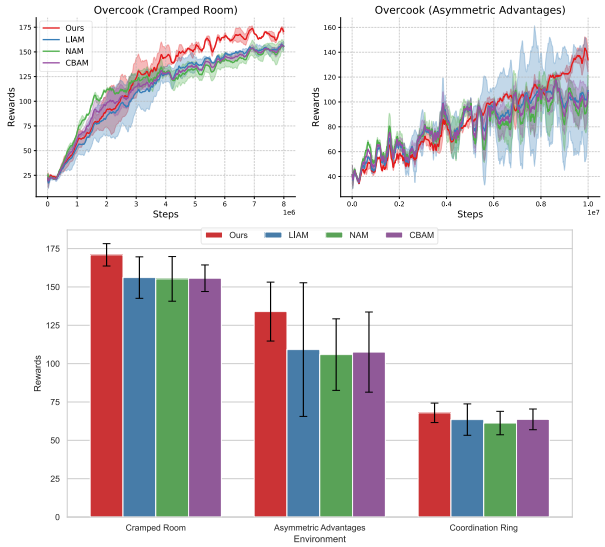

Figure 5: Performance Against Human Proxies. This figure shows the performance of our method compared to the baseline in human-AI interaction settings. Our method consistently achieves better results than the baseline.

zero-shot coordination approaches, which lack opponent modeling. Additional results are in Appendix G.

5.2.2 Performance. The comparison between our method and the baseline in the Overcooked environment is presented in Figure 5. The experimental results demonstrate the consistently high performance of our approach in human-AI interaction tasks. Methods that do not account for uncertainty, such as LIAM, struggle with complex opponent behaviors in more challenging tasks like Asymmetric Advantages, resulting in significant performance variance. In contrast, our method exhibits greater stability across all tasks and continues to improve steadily throughout the training process.

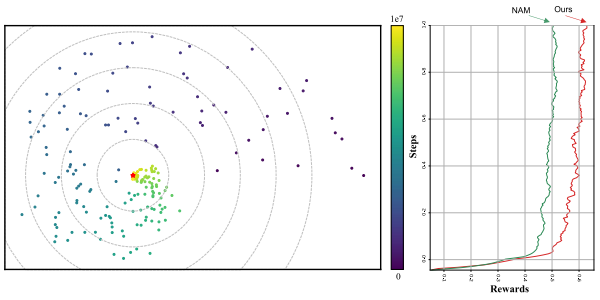

Figure 6: t-SNE Visualization of Opponent Embeddings Over Time. This figure shows the evolution of opponent embeddings learned by our method over time. The embeddings stabilize and converge closely to the true values by the end of the training. This visualization reflects both the convergence and accuracy of our opponent features, indicating that our method effectively handles aleatoric uncertainty.

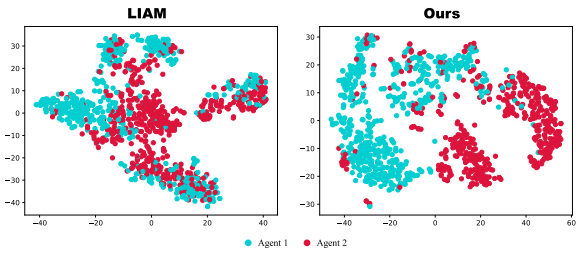

Figure 7: Comparison of Opponent Embeddings via t-SNE. This figure compares t-SNE mappings of opponent embeddings. Compared to methods that do not account for uncertainty (e.g., LIAM), our approach demonstrates superior discriminability of opponents, indicating that our model better addresses epistemic uncertainty.

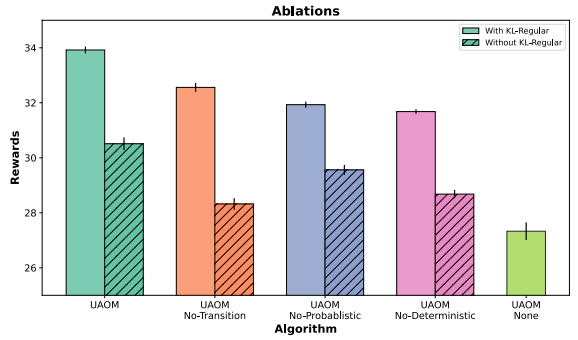

Figure 8: Ablation Study of UAOM. This figure highlights that both the probabilistic and deterministic networks are essential, with each playing a crucial role. The transition network further improves performance, and KL divergence helps maintain stability across the modules.

5.3 Analysis and Ablations

To evaluate the effectiveness of our method in handling uncertainty in opponent modeling, we examined three key aspects: discriminability, convergence, and accuracy. High discriminability indicates that the model effectively addresses epistemic uncertainty stemming from its own cognitive limitations [30]. Stable convergence demonstrates that the model successfully mitigates aleatoric uncertainty arising from data randomness [3]. Finally, accuracy reflects the model's ability to handle uncertainty from multiple sources [2]. Our experimental results show that our method performs well in all three aspects.

We first assessed the discriminability of our opponent modeling approach. We visualized the opponent embeddings in the Level Based Foraging (Large) environment with a regularization parameter $\lambda = 2$. We employed t-SNE to reduce the high-dimensional feature space into two dimensions, allowing for a more intuitive understanding of the clustering behavior. As shown in Figure 6, the embeddings for different opponent strategies are displayed in two plots. In the right plot, which corresponds to the LIAM method, the embeddings are visibly overlapping, indicating poor separation between the distinct opponent strategies. In contrast, the left plot, representing our proposed method, shows well-separated clusters of embeddings. This clear distinction between the clusters highlights that our approach is more effective in capturing and differentiating the unique behavioral traits of opponent behaviors. The separation of the embeddings demonstrates a significant enhancement in dealing with epistemic uncertainty as compared to LIAM, which does not account for such uncertainty.

We evaluated the convergence and accuracy of our opponent modeling method by analyzing the temporal evolution of opponent embeddings, as shown in Figure 7. We first characterized a deterministic opponent ($\lambda = \infty$) as ground truth and then examined a stochastic opponent ($\lambda = 2$) using our approach. Over time, the embeddings for the stochastic opponent converged to stable values close to the ground truth, indicating our method effectively captures the underlying structure of opponent behavior despite randomness. Furthermore, the learning process of embeddings strongly correlated with improved performance in the reinforcement learning task, highlighting our method's ability to capture behavioral nuances for better decision-making. The analysis of both convergence and accuracy for high-randomness opponents further confirms that our method effectively mitigates the impact of uncertainty. For a more detailed numerical analysis of accuracy, please refer to Appendix H.

Additionally, we conducted ablation studies to assess the contribution of each module in our method. As shown in Figure 8, the ablation results confirm that all components of our approach are effective. Both the probabilistic and deterministic networks play crucial roles, with neither being dispensable. The transition network further enhances performance, while the KL divergence term helps maintain the stability of the model across different modules.

6 CONCLUSION

We propose the UAOM method, which effectively addresses uncertainty in opponent modeling through its hybrid behavior modeling and opponent-aware dynamic modeling components. Experimental results demonstrate the consistent effectiveness of our approach in both simulation settings and human-agent interactions. Our method tackles the prevalent uncertainty issues in opponent modeling and provides insights for handling human opponents in practical applications. In future research, we plan to expand our approach to multi-agent reinforcement learning, focusing on how opponent modeling can be effectively adapted to scenarios with high uncertainty, where multiple agents learn and interact concurrently. This extension could provide deeper insights into the dynamics of complex environments with numerous interacting entities.

ACKNOWLEDGMENTS

This work is supported in part by National Science and Technology Major Project, Grant No.2022ZD0116403, the National Natural Science Foundation of China, Grant No.62176255, the Postdoctoral Fellowship Program of CPSF, Grant No.GZC20232995, and the China Postdoctoral Science Foundation, Grant No. 2024M763533.

REFERENCES

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

[2] Javier Antorán, James Allingham, and José Miguel Hernández-Lobato. 2020. Depth uncertainty in neural networks. *Advances in neural information processing systems* 33 (2020), 10620–10634.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*. PMLR, 1613–1622.

[4] Roberto Calandra, Jan Peters, Carl Edward Rasmussen, and Marc Peter Deisenroth. 2016. Manifold Gaussian processes for regression. In *2016 International joint conference on neural networks (IJCNN)*. IEEE, 3338–3345.

[5] Colin F Camerer, Teck-Hua Ho, , and Juin-Kuan Chong. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics* 119, 3 (2004), 861–898.

[6] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems* 32 (2019).

[7] Yali Du, Joel Z Leibo, Usman Islam, Richard Willis, and Peter Sunehag. 2023. A review of cooperation in multi-agent learning. *arXiv preprint arXiv:2312.05162* (2023).

[8] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779* (2016).

[9] Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2017. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326* (2017).

[10] Eric A Hansen, Daniel S Bernstein, , and Shlomo Zilberstein. 2004. Dynamic programming for partially observable stochastic games. In *AAAI*, Vol. 4. 709–715.

[11] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In *International conference on machine learning*. PMLR, 1804–1813.

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

[13] Eric Jang, Shixiang Gu, , and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[14] Yuheng Jing, Kai Li, Bingyun Liu, Yifan Zang, Haobo Fu, QIANG FU, Junliang Xing, and Jian Cheng. 2023. Towards Offline Opponent Modeling with In-context Learning. In *The Twelfth International Conference on Learning Representations*.

[15] Akbir Khan, Timon Willi, Newton Kwan, Andrea Tacchetti, Chris Lu, Edward Grefenstette, Tim Rocktäschel, and Jakob Foerster. 2023. Scaling opponent shaping to high dimensional games. *arXiv preprint arXiv:2312.12568* (2023).

[16] Durk P Kingma, Tim Salimans, , and Max Welling. 2015. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems* 28 (2015).

[17] Balaji Lakshminarayanan, Alexander Pritzel, , and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[18] Christopher Lu, Timon Willi, Christian A Schroeder De Witt, and Jakob Foerster. 2022. Model-free opponent shaping. In *International Conference on Machine Learning*. PMLR, 14398–14411.

[19] Richard D McKelvey and Thomas R Palfrey. 1995. Quantal response equilibria for normal form games. *Games and economic behavior* 10, 1 (1995), 6–38.

[20] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.

[21] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[22] Samer Nashed and Shlomo Zilberstein. 2022. A survey of opponent modeling in adversarial domains. *Journal of Artificial Intelligence Research* 73 (2022), 277–327.

[23] Martin A Nowak. 2006. Five rules for the evolution of cooperation. *science* 314, 5805 (2006), 1560–1563.

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

[25] Georgios Papoudakis, Filippos Christianos, , and Stefano Albrecht. 2021. Agent modelling under partial observability for deep reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 19210–19222.

[26] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*. PMLR, 4218–4227.

[27] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. 2018. Modeling others using oneself in multi-agent reinforcement learning. In *International conference on machine learning*. PMLR, 4257–4266.

[28] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.

[29] Paul Slovic, Baruch Fischhoff, , and Sarah Lichtenstein. 1977. Behavioral decision theory. *Annual review of psychology* 28, 1 (1977), 1–39.

[30] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*. PMLR, 9690–9700.

[31] Ying Wen, Yaodong Yang, Rui Luo, and Jun Wang. 2019. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. *arXiv preprint arXiv:1901.09216* (2019).

[32] Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. 2019. Probabilistic recursive reasoning for multi-agent reinforcement learning. *arXiv preprint arXiv:1901.09207* (2019).

[33] Pei Xu, Junge Zhang, and Kaiqi Huang. 2024. Population-Based Diverse Exploration for Sparse-Reward Multi-Agent Tasks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. 283–291. https://doi.org/10.24963/ijcai.2024/32

[34] Pei Xu, Junge Zhang, Qiyue Yin, Chao Yu, Yaodong Yang, and Kaiqi Huang. 2023. Subspace-Aware Exploration for Sparse-Reward Multi-Agent Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI-23*. 11717–11725. https://doi.org/10.1609/aaai.v37i10.26384

[35] Xiaopeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. 2022. Model-based opponent modeling. *Advances in Neural Information Processing Systems* 35 (2022), 28208–28221.

[36] Lei Yuan, Ziqian Zhang, Lihe Li, Cong Guan, and Yang Yu. 2023. A survey of progress on cooperative multi-agent reinforcement learning in open environment. *arXiv preprint arXiv:2312.01058* (2023).

[37] Stephen Zhao, Chris Lu, Roger B Grosse, and Jakob Foerster. 2022. Proximal Learning With Opponent-Learning Awareness. *Advances in Neural Information Processing Systems* 35 (2022), 26324–26336.

[38] Luisa Zintgraf, Sebastian Schulze, Cong Lu, Leo Feng, Maximilian Igl, Kyriacos Shiarlis, Yarin Gal, Katja Hofmann, and Shimon Whiteson. 2021. Varibad: Variational bayes-adaptive deep rl via meta-learning. *The Journal of Machine Learning Research* 22, 1 (2021), 13198–13236.