Innate-Values-driven Reinforcement Learning for Cooperative Multi-Agent Systems

Qin Yang[1*]

[1]Computer Science and Information Systems Department, Bradley University
1501 W Bradley Ave, Bradley Hall 195, Peoria, IL 61625, USA
Email: is3rlab@gmail.com

Abstract

Innate values describe agents' intrinsic motivations, which reflect their inherent interests and preferences to pursue goals and drive them to develop diverse skills satisfying their various needs. The essence of reinforcement learning (RL) is learning from interaction based on reward-driven (such as utilities) behaviors, much like natural agents. It is an excellent model to describe the innate-values-driven (IV) behaviors of AI agents. Especially in multi-agent systems (MAS), building the awareness of AI agents to balance the group utilities and system costs and satisfy group members' needs in their cooperation is a crucial problem for individuals learning to support their community and integrate human society in the long term. This paper proposes a hierarchical compound intrinsic value reinforcement learning model – innate-values-driven reinforcement learning termed IVRL to describe the complex behaviors of multi-agent interaction in their cooperation. We implement the IVRL architecture in the StarCraft Multi-Agent Challenge (SMAC) environment and compare the cooperative performance within three characteristics of innate value agents (Coward, Neutral, and Reckless) through three benchmark multi-agent RL algorithms: QMIX, IQL, and QTRAN. The results demonstrate that by organizing individual various needs rationally, the group can achieve better performance with lower costs effectively.

Introduction

In natural systems, motivation is concerned explicitly with the activities of creatures that reflect the pursuit of a particular goal and form a meaningful unit of behavior in this function (**?**). Furthermore, intrinsic motivations describe incentives relating to an activity itself, and these incentives residing in pursuing an activity are intrinsic. Intrinsic motivations deriving from an activity may be driven primarily by interest or activity-specific incentives, depending on whether the object of an activity or its performance provides the main incentive (**?**). They also fall in the category of cognitive motivation theories, which include theories of the mind that tend to be abstracted from the biological system of the behaving organism (**?**).

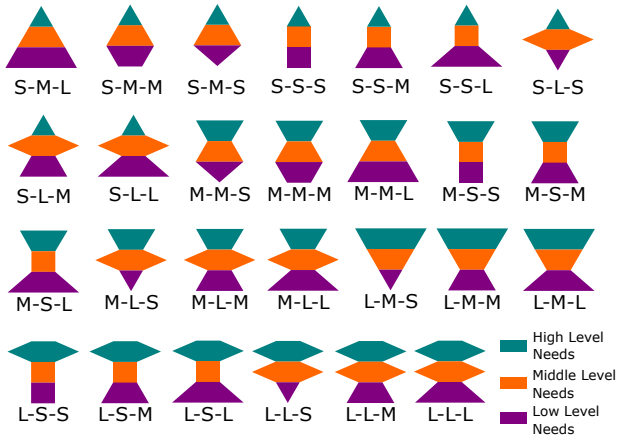


Figure 1: The illustration of innate values models with three-level needs of different amounts. *S: Small; M: Medium; L: Large*

However, when we analyze natural agents, such as humans, they are usually combined motivation entities. They have biological motivations, including physiological, safety, and existence needs; social motivation, such as love and esteem needs; and cognitive motivation, like self-actualization or relatedness and growth needs (**?**). The combined motivation theories include Maslow's Hierarchy of Needs (**?**) and Alderfer's Existence Relatedness Growth (ERG) theory (**?**). Fig. 1 illustrates innate values (intrinsic motivations) models with three-level needs of different amounts.

Many researchers regard motivated behavior as behavior that involves the assessment of the consequences of behavior through learned expectations, which makes motivation theories tend to be intimately linked to theories of learning and decision-making (**?**). In particular, intrinsic motivation leads organisms to engage in exploration, play, strategies, and skills driven by expected rewards. The computational theory of reinforcement learning (RL) addresses how predictive values can be learned and used to direct behavior, making RL naturally relevant to studying motivation.

In artificial intelligence, researchers propose various ab-

[*]Qin Yang is the director of the Intelligent Social Systems and Swarm Robotics Lab (IS[3]R), Computer Science and Information Systems Department, Bradley University, Peoria, IL 61625, USA; Email: is3rlab@gmail.com.

stract computational structures to form the fundamental units of cognition and motivations, such as states, goals, actions, and strategies. For intrinsic motivation modeling, the approaches can be generally classified into three categories: prediction-based (**??**), novelty-based (**??**), and competence-based (**??**). Furthermore, the concept of intrinsic motivation was introduced in machine learning and robotics to develop artificial systems learning diverse skills autonomously. The idea is that intelligent machines and robots could autonomously acquire skills and knowledge under the guidance of intrinsic motivations and later exploit such knowledge and skills to accomplish tasks more efficiently and faster than if they had to acquire them from scratch (**?**).

In other words, by investigating intrinsically motivated learning systems, we would clearly improve the utility and autonomy of intelligent artificial systems in dynamic, complex, and dangerous environments (**??**). Specifically, compared with the traditional RL model, intrinsically motivated RL refines it by dividing the environment into an external environment and an internal environment, which clearly generates all reward signals within the organism [1] (**?**). However, although the extrinsic reward signals are triggered by the objects and events of the external environment, and activities of the internal environment cause the intrinsic reward signals, it is hard to determine the complexity and variability of the intrinsic rewards (innate values) generating mechanism.

From the MAS perspective, learning in multi-agent settings is fundamentally more difficult than in the single-agent case due to the presence of multi-agent pathologies, e.g., the moving target problem (non-stationarity), curse of dimensionality, multi-agent credit assignment, global exploration, and relative overgeneralization (**?**). Furthermore, due to each agent facing a moving-target learning issue, the best policy might change as the other agents modify their policies since all the agents are learning simultaneously in the process, which raises the nonstationarity of the multi-agent learning problem (**?**). Therefore, multi-agent RL algorithms need to balance exploiting the agent's current knowledge and exploring information-gathering actions to improve its knowledge. Especially for the cooperative MAS setting, individual choices of actions and strategies should be mutually consistent in order to achieve their common goals.

Moreover, most MAS implementations aim to optimize the system's policies with respect to individual needs and intrinsic values, even though many real-world problems are inherently multi-objective (**?**). Thus, many conflicts and complex trade-offs in the MAS need to be managed, and compromises among agents should be based on the utility mapping the innate values of a compromise solution – how to measure and what to optimize (**?**). However, in the MAS setting, the situations will become much more complex when we consider individual utility reflects its own needs and preferences (**??**). For example, although we assume each group member receives the same team rewards in fully cooperative MAS, the benefits received by an individual agent are usually significant differences according to its contributions and

[1]Here, the organism represents all the components of the internal environment in the AI agent.

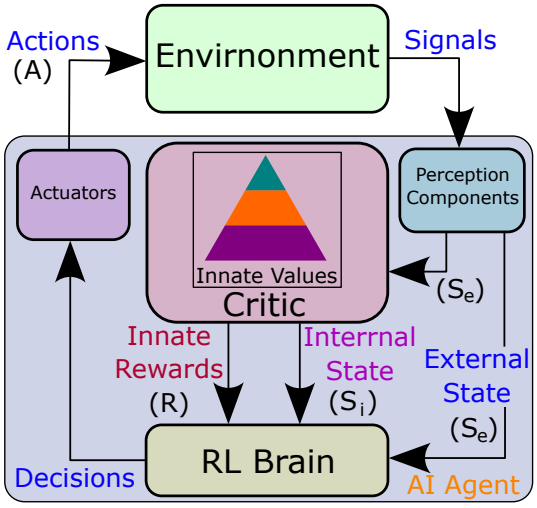


Figure 2: The illustration of the proposed innate-values-driven reinforcement learning (IVRL) model.

innate values in real-world scenarios or general multi-agent settings (**?**).

To address those gaps, through integrating with combined motivation theories, we propose a novel innate-values-driven reinforcement learning (IVRL) model to describe the complex behaviors in MAS cooperation. By demonstrating the proposed architecture in the SAMC environment with three benchmark multi-agent RL algorithms (QMIX, IQL, and QTRAN), we prove that rationally organizing various individual needs can achieve better performance with lower costs in MAS cooperation effectively.

Approach Overview

We assume that all the AI agents (like robots) interact in the same working scenario, and their external environment includes all the other group members and mission setting. In contrast, the internal environment consists of individual perception components including various sensors (such as Lidar and camera), the critic module involving intrinsic motivation analysis and innate values generation, the RL brain making the decision based on the feedback of rewards and description of the current state (including internal and external) from the critic module, and actuators relating to all the manipulators and operators executing the RL brain's decisions as action sequence and strategies. Fig. 2 illustrates the proposed IVRL model.

Compared with the traditional RL model, in our model, the input state and rewards are generated from the critic module instead of directly coming from the environment, which means that the individual needs to calculate benefits or utilities based on the innate value model and update its current internal status (the amount of the current needs locating in different levels) combining with external state sending to the RL model. For example, supposing two agents a^1 and a^2 have different innate value models: Fig. 1 S-M-L and L-M-S. We use health points, energy levels, and task achieve-

ment to represent their safety needs n_s (low-level intrinsic motivations), basic needs n_b (middle-level intrinsic motivations), and teaming needs n_t (High-level intrinsic motivations) (**???**), respectively. Considering $n_s^1 > n_s^2$, $n_b^1 = n_b^2$, and $n_t^1 < n_t^2$, if a^1 and a^2 receive the external repairing signal simultaneously, the innate rewards r^1 will larger than r^2 based on their innate value model (low-level safety needs $n_s^1 > n_s^2$). In contrast, if they get the order to collaborate with other agents fulfilling a task (high-level teaming needs $n_t^1 < n_t^2$) at the same time, the agent a^2 will receive more credits from it ($r^1 < r^2$). In other words, due to different innate value models, their intrinsic motivations present significant differences, which will lead to different innate rewards for performing the same task. Moreover, after calculating the practical credits for the individual utility in current actions, it will update its intrinsic status and combine with the external environment state sending to the RL model.

Furthermore, in the long term, the individual innate value model is dynamically changed after various needs at different levels are satisfied in the multi-agent interaction. It means that the agent's innate reward mechanism is not fixed, which helps it develop diverse skills and strategies to handle complex, multi-object, dynamic, and uncertain environments in the mission, similar to humans. Moreover, we formalize the proposed IVRL model as below.

Single-Agent Systems

In the single-agent systems, we formalize the IVRL of an AI agent with an external environment using a Markov decision process (MDP) (**?**). The MDP is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ where $\mathcal{S}$ represents the finite sets of internal state S_i[2] and external states S_e. $\mathcal{A}$ represents a finite set of actions. The transition function $\mathcal{T}$: $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow$ [0, 1] determines the probability of a transition from any state $s \in \mathcal{S}$ to any state $s' \in \mathcal{S}$ given any possible action $a \in \mathcal{A}$. Assuming the critic function is $\mathcal{C}$, which describes the individual innate value model. The reward function $\mathcal{R} = \mathcal{C}(S_e) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ defines the immediate and possibly stochastic innate reward $\mathcal{C}(S_e)$ that an agent would receive given that the agent executes action a which in state s and it is transitioned to state $s', \gamma \in [0, 1)$ the discount factor that balances the trade-off between innate immediate and future rewards (Fig. 2).

The IVRL are adequate model to obtain optimal decisions in single agent fully observable environments. Solving the IVRL model will yield a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, which is a mapping from internal and external state to actions. An optimal policy π^* is the one that maximizes the expected discounted sum of innate rewards Eq. (1).

$$\mathbb{E}\left[\sum_{t\geq 0} \gamma^t \mathcal{C}(s_{i_t}, a_t, s_{i_{t+1}}) \middle| a_t \sim \pi(\cdot|s_t), s_0\right] \tag{1}$$

Accordingly, we can define the *action-innate-value* function (Q-function) and *state-innate-value* function (V-function) under policy π as Eq. (2) and (3) for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, which are the discounted accumulated innate reward starting from $(s_0, a_0) = (s, a)$ and $s_0 = s$, respectively.

$$\mathbf{Q}_\pi(s, a) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \mathcal{C}(s_{i_t}, a_t, s_{i_{t+1}}) \middle| a_t \sim \pi(\cdot|s_t), a_0 = a, s_0 = s\right] \tag{2}$$

$$\mathbf{V}_\pi(s, a) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \mathcal{C}(s_{i_t}, a_t, s_{i_{t+1}}) \middle| a_t \sim \pi(\cdot|s_t), s_0 = s\right] \tag{3}$$

Here, the ones corresponding to the optimal policy π^* are referred to as the optimal Q-function and the optimal state-innate-value function, respectively.

Multi-Agent Systems

In the multi-agent setting, we consider the Markov games, also known as stochastic games (**?**), originating from the seminal work (**?**). We formalize the IVRL of the MAS as below:

A Markov game is defined by a tuple $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \{\mathcal{R}^i\}_{i\in\mathcal{N}}, \gamma)$, where $\mathcal{N} = \{1, \ldots, N\}$ denotes the set of $N > 1$ agents, $\mathcal{A}^i$ denotes the action space of agent i. Let $\mathcal{A} := \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, then $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ denotes the transition probability from any state $s \in \mathcal{S}$ to any state $s' \in \mathcal{S}$ (including internal state $\mathcal{S}_i$ and external state $\mathcal{S}_e$) for any joint action $a \in \mathcal{A}$. For the agent i, $\mathcal{C}^i$ is the critic function of the innate value model. $\mathcal{R}^i = \mathcal{C}^i(\mathcal{S}_e^i) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the innate reward function that determines the immediate reward received by agent i for a transition from (s, a) to s'. $\gamma \in [0, 1)$ is the discount factor.

Any time t, each agent $i \in \mathcal{N}$ executes an action a_t^i, according to the system state s_t. The system then transitions to state s_{t+1}, and innate rewards of each agent i by $\mathcal{R}^i = \mathcal{C}^i(s_{i_t}, a_t, s_{i_{t+1}})$. The goal of agent i is to optimize its own long-term innate reward, by finding policy $\pi^i : \mathcal{S} \rightarrow \Delta(\mathcal{A}^i)$ such that $a_t^i \dim \pi^i(\cdot|s_t)$. As a consequence, the innate value function $V^i : \mathcal{S}_i \rightarrow \mathbb{R}$ of agent i becomes a function of joint policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ defined as $\pi(a|s) := \prod_{i\in\mathcal{N}} \pi^i(a^i|s)$. In particular, for any joint policy π and state $s \in \mathcal{S}$, we can get Eq. (4), where $-i$ represents the indices of all agents in $\mathcal{N}$ except agent i.

$$\mathbf{V}^i_{\pi^i, \pi^{-i}}(s) := \mathbb{E}\left[\sum_{t\geq 0} \gamma^t \mathcal{C}^i(s_{i_t}, a_t, s_{i_{t+1}}) \middle| a_t^i \sim \pi^i(\cdot|s_t), s_0 = s\right] \tag{4}$$

Here, the optimal performance of each agent is controlled not only by its own policy but also by the choice of all other group members of the game.

[2]The internal state S_i describes an agent's innate value distribution and presents the dominant intrinsic motivation based on the external state S_e.

Cooperative Setting

In the IVRL model, the innate value function describes the agent's current motivation status and dominant motivation. If agents have similar distributions of needs or motivations, they will share a common innate value function, i.e., $\mathcal{C}^1 = \mathcal{C}^2 = \cdots = \mathcal{C}^{\mathcal{N}}$, which means that they have the potential to collaborate in the current situation (**?**). Moreover, from the game-theoretic perspective, if all agents' innate value functions and Q-functions are identical, they can be coordinated as one decision-maker, and the global optimum for cooperation now constitutes a Nash equilibrium of the game.

However, the innate value distributions of agents will dynamically change (Fig. 1) due to satisfying individuals' current needs and motivations. In the long term, in order to optimize the global system's utility and guarantee sustainable development for each group member, we can dynamically group the agents of similar innate value models and assign them corresponding tasks. Moreover, we can optimize the average innate reward $\bar{\mathcal{C}}(s_e, a, s'_e) := \mathcal{N}^{-1} \cdot \sum_{i \in \mathcal{N}} \mathcal{C}^i(s_e, a, s'_e)$, which allow to have different innate reward functions and more heterogeneity among agents (agents with different motivations and needs), similar to human society.

Evaluation through Simulation Studies

We evaluate the performance of the proposed innate-value-driven reinforcement learning (IVRL) model in the StarCraft Multi-Agent Challenge (SMAC) (**?**) environment. Our experiments define three types of group agents based on the proposed innate-value model[3], which presents corresponding group personalities: *Coward*, *Neutral*, and *Reckless*. Furthermore, we study the group performance of different characteristics in several benchmark multi-agent RL algorithms, such as QMIX (**?**), IQL (**?**), and QTRAN (**?**). And we discuss more details as below.

Experiment Setting

In our experiment, we use a weight matrix to model the reward mechanism of the innate value system. To simplify the computational process, we only consider three factors – battle won (BW), shield level (SL), and health points (HP) – (Eq. (5)) describing the innate value distribution of an agent.

$$\mathbb{W}_{iv} = \begin{bmatrix} \mathbf{Achievement}_{high\_level\_needs} \\ \mathbf{Safety}_{middle\_level\_needs} \\ \mathbf{Basic}_{low\_level\_needs} \end{bmatrix} = \begin{bmatrix} \mathcal{BW}_{weight} \\ \mathcal{SL}_{weight} \\ \mathcal{HP}_{weight} \end{bmatrix} \tag{5}$$

Specifically, according to the Eq. (5), we define the agent of the coward, neutral, and reckless based on different innate value weights in the matrix, which express distinguished rewards with the same actions or strategies when it interacts with the environments and other agents. For example, the coward cares much about its life and HP. On the other hand, the reckless only eyes on the result of the battle won regardless of its life and safety. Compared with the two types of agents, the neutral agent represents more rationality to balance the low-level basic and safety needs and high-level teaming needs (battle won). Moreover, we can formalize the coward (Eq. (6)), neutral (Eq. (7)), and reckless (Eq. (8)) as follow:

$$\mathcal{BW}_{weight} \ll \mathcal{SL}_{weight} \approx \mathcal{HP}_{weight}; \tag{6}$$

$$\mathcal{BW}_{weight} \approx \mathcal{SL}_{weight} \approx \mathcal{HP}_{weight}; \tag{7}$$

$$\mathcal{BW}_{weight} \gg \mathcal{SL}_{weight} \approx \mathcal{HP}_{weight}; \tag{8}$$

Considering integrating the agent's parameters of the StatCraft II into the individual innate-value system, we also use the corresponding health points and shield level representing its basic and safety needs, respectively. Moreover, we apply the result of the battle won for each episode to describe the teaming goal as the individual high-level needs achieving in their cooperation. We implement three benchmark multi-agent RL algorithms, such as QMIX, IQL, and QTRAN, in the standard StarCraft map 2s3z to evaluate the battle performance of the three types group of agents (Eq. (6), (7), and (8)). Furthermore, we list the innate-value weight matrix of each type of agent in corresponding experiments as Tab. 1. Fig. 3 illustrates the 2s3z testing Map in the StarCraft II Multi-Agent Challenge (SMAC) Experiment Environment.

Table 1: Innate-Value Weight Matrix of Each Type of Agent in Experiments.

Weight Matrix / CHAR / ALG	Coward	Neutral	Reckless
QMIX	$[1, -2.5, -2.5]^T$	$[1, -1, -1]^T$	$[1, 2.5, 2.5]^T$
IQL	$[1, -2.5, -2.5]^T$	$[1, -1, -1]^T$	$[1, 2.5, 2.5]^T$
QTRAN	$[1, -3, -3]^T$	$[1, -1, -1]^T$	$[1, 3, 3]^T$

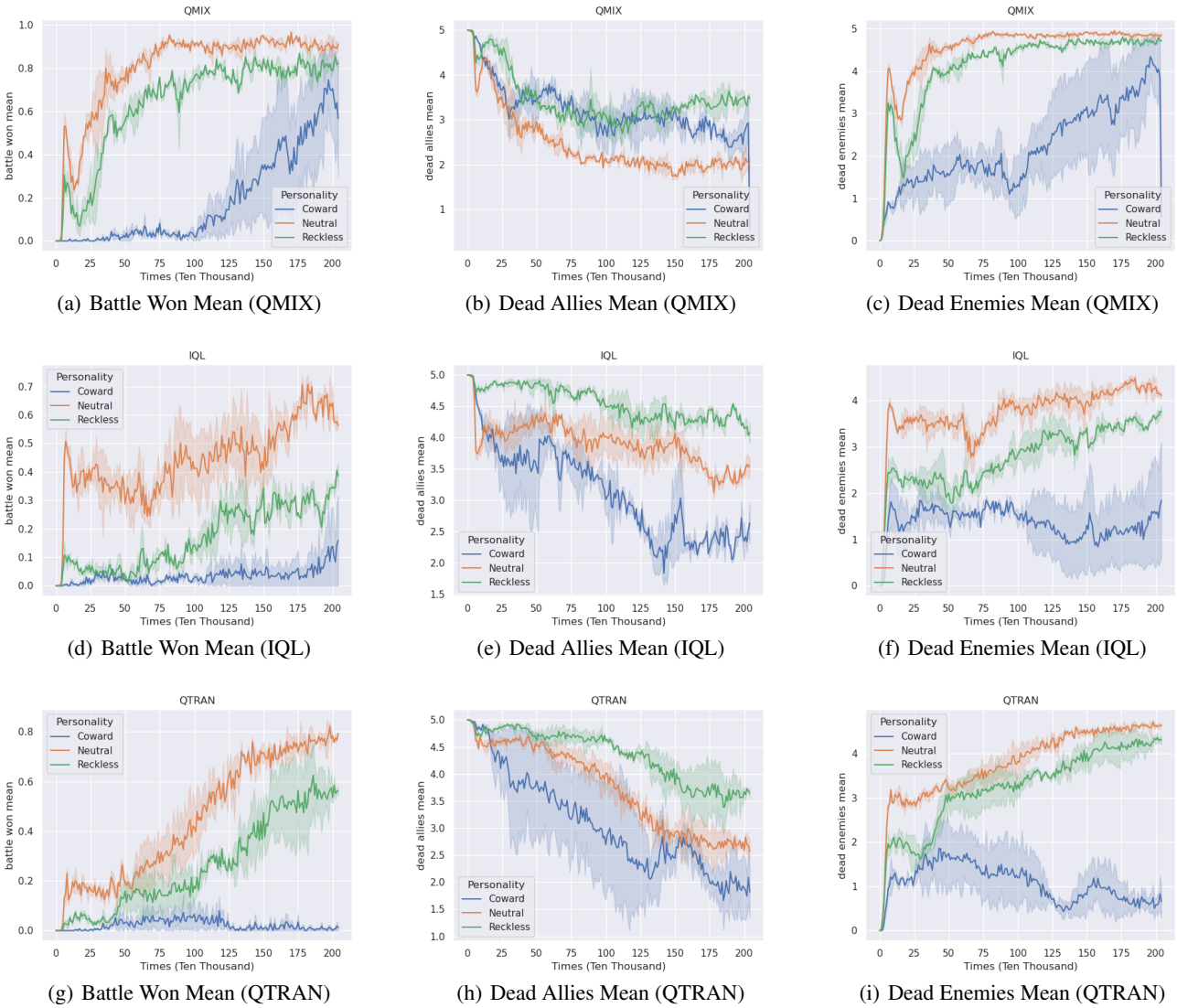
Generally, different characteristic agents present unique innate-value systems corresponding to the specific reward mechanism to interact with other agents and various environments.

Evaluation and Results

We first study the proposed IVRL architecture with the QMIX algorithm. Fig. 4(a) shows that the group of neutral agents has the highest battle-won mean compared with the reckless group and coward group. For the innate-value system of cowards, since they care more about low-level needs, like basic and safety needs, their personality presents a kind of self-interest in the battle. It means selfishness and greedy behaviors can achieve more rewards in their value systems. Therefore, these individual behaviors significantly damage the benefits of the entire group and hinder the team's cooperation to fulfill the task and reach high-level goals. Moreover, since group benefits have been seriously damaged, it has impacted individual interests substantially. Fig. 4(b) and 4(c) demonstrate that the coward group loses more allies but destroys fewer enemies.

[3]This experiment only consider static innate-value system setting.

Figure 3: Illustration of the 2s3z Map in the StarCraft II Multi-Agent Challenge (SMAC) Experiment Environment



(a) Battle Won Mean (QMIX) (b) Dead Allies Mean (QMIX) (c) Dead Enemies Mean (QMIX)

(d) Battle Won Mean (IQL) (e) Dead Allies Mean (IQL) (f) Dead Enemies Mean (IQL)

(g) Battle Won Mean (QTRAN) (h) Dead Allies Mean (QTRAN) (i) Dead Enemies Mean (QTRAN)

Figure 4: The group performance of three personalities (innate-value) agents with QMIX, IQL, and QTRAN in the SMAC

needs (Fig. 4(b) and 4(c)). It also affects the group to achieve better performance (Fig. 4(a)). Furthermore, the neutral group represents more rational behaviors in the battle, balancing the team's and individual interests. In other words, based on the reward mechanism of the innate value system, the neutral agents optimize the global system's utility and guarantee sustainable development for each group member by balancing the rewards between agents and groups in MAS cooperation, much like human society does.

in the IVRL architecture. They all show similar results as the QMIX (Fig. 4(d) and 4(g)). Especially in the Fig. 4(e) and 4(h), they illustrate that the coward group's behaviors can save more allies' lives, but the reckless group's behaviors cost a lot.

various needs in the systems. It also builds different personalities and characteristics of agents in their interaction. Moreover, organizing agents with similar interests and innate values in the mission can optimize the group utilities and reduce costs effectively, just like "Birds of a feather flock together." in human society.

Conclusion and Future Works

different scenarios, such as single-agent, multi-agent, and cooperative settings. Furthermore, we demonstrate the proposed architecture in the SAMC environment with three benchmark multi-agent RL algorithms (QMIX, IQL, and QTRAN). The results prove that rationally organizing various individual needs can achieve better performance with lower costs in MAS cooperation effectively.

testbeds, such as StarCraft II, OpenAI Gym, Unity, etc. Moreover, integrating efficient deep RL algorithms, such as BSAC (**??**), with the IVRL can help agents evolve diverse skills to adapt to complex environments in MAS cooperation. Furthermore, implementing the IVRL in real-world systems, such as human-robot interaction, multi-robot systems, and self-driving cars, would be challenging and exciting.