frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 06 July 2023
DOI 10.3389/fnins.2023.1219405

Check for updates

OPEN ACCESS

EDITED BY
Malu Zhang,
National University of Singapore, Singapore

REVIEWED BY
Xiurui Xie,
University of Electronic Science and Technology of China, China
Pengfei Sun,
Ghent University, Belgium

*CORRESPONDENCE
Tielin Zhang
✉ tielin.zhang@ia.ac.cn
Bo Xu
✉ xubo@ia.ac.cn

RECEIVED 09 May 2023
ACCEPTED 07 June 2023
PUBLISHED 06 July 2023

CITATION
Li X, Ni Z, Ruan J, Meng L, Shi J, Zhang T and Xu B (2023) Mixture of personality improved spiking actor network for efficient multi-agent cooperation. *Front. Neurosci.* 17:1219405. doi: 10.3389/fnins.2023.1219405

COPYRIGHT
© 2023 Li, Ni, Ruan, Meng, Shi, Zhang and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Mixture of personality improved spiking actor network for efficient multi-agent cooperation

Xiyun Li[1,2], Ziyi Ni[1,3], Jingqing Ruan[1,2], Linghui Meng[1,3], Jing Shi[1,3], Tielin Zhang[1,3]* and Bo Xu[1,2,3,4]*

[1]Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China, [3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [4]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

Adaptive multi-agent cooperation with especially unseen partners is becoming more challenging in multi-agent reinforcement learning (MARL) research, whereby conventional deep-learning-based algorithms suffer from the poor new-player-generalization problem, possibly caused by not considering theory-of-mind theory (ToM). Inspired by the ToM personality in cognitive psychology, where a human can easily resolve this problem by predicting others' intuitive personality first before complex actions, we propose a biologically-plausible algorithm named the mixture of personality (MoP) improved spiking actor network (SAN). The MoP module contains a determinantal point process to simulate the formation and integration of different personality types, and the SAN module contains spiking neurons for efficient reinforcement learning. The experimental results on the benchmark cooperative overcooked task showed that the proposed MoP-SAN algorithm could achieve higher performance for the paradigms with (learning) and without (generalization) unseen partners. Furthermore, ablation experiments highlighted the contribution of MoP in SAN learning, and some visualization analysis explained why the proposed algorithm is superior to some counterpart deep actor networks.

KEYWORDS

multi-agent cooperation, personality theory, spiking actor networks, multi-agent reinforcement learning, theory of mind

1. Introduction

With the rapid development and great progress of deep reinforcement learning (RL) in recent years (Silver et al., 2017; Vaswani et al., 2017; Vinyals et al., 2019; Yu et al., 2021; Meng et al., 2023), more and more researchers have shown an increased interest in multi-agent cooperation or human-in-the-loop cooperation (Carroll et al., 2019; Shih et al., 2021, 2022; Strouse et al., 2021; Zhao et al., 2021; Ruan et al., 2022; Lou et al., 2023). However, cooperation with unseen partners usually requires continuous collection of expert data, which is expensive and delayed (Carroll et al., 2019; Shih et al., 2022). Other methods

attempt to achieve better generalization without expert data by constructing a population pool for simulating diverse candidate partners. However, these studies try to improve the generalization cooperation score by relying on being trained with a large number of well-designed partners but ignore the cultivation of the agent's real thinking and empathy ability.

The less consideration of the psychological characteristics of partner agents might be the key reason why these artificial agents fail, compared to their counterpart biological agents. In our daily life, humans can cooperate well with others whom we have never seen before (Boyd and Richerson, 2009; Rand and Nowak, 2013). This phenomenon is interesting but not hard to guess. We can infer others' personalities quickly, and then we can well handle the following cooperation behaviors with the help of this guessed personality. The personality theory is under the framework of theory of mind (ToM) (Gallagher and Frith, 2003; Frith and Frith, 2005; Roth et al., 2022; Aru et al., 2023), which refers to our ability to speculate on the intentions, behaviors, and goals of other people, which explains why humans can collaborate with unseen partners from a cognitive perspective. In fact, instead of being classified into a specific personality, the unseen human can be viewed as some combination of several "personalities." Therefore, it is significantly helpful to find as few representative personalities as possible and make them orthogonal to each other for a more efficient combination. The personality theory (McCrae and Costa, 2008; Ryckman, 2012; Schultz and Schultz, 2016) from cognitive psychology has provided an opportunity to model the partners more clearly and concretely, including the big five personalities (De Raad, 2000) and the sixteen personality factors (16PF) (Cattell and Mead, 2008). These theories are useful in describing unique and diverse people (Anglim and Horwood, 2021) and can instruct many cognitive tasks, such as personality trait tests (O'Connor and Paunonen, 2007) to analyze people's suitable careers.

Unlike the personality theory in cognitive science, which is often used as the discrete classification, we propose the base personality similar to the base vector in the personality space, which can be used for inferring personality. To further ensure the difference between multiple base personalities, determinantal point process (DPP) constraints are adopted as an intrinsic reward. Based on the personality model with these base personalities, the agent can naturally predict and understand any unseen partner to better make responses and obtain cooperation.

Hence, inspired by the above personality theory, we propose the mixture of personality (MoP), along with our previously proposed spiking agent network (SAN), which has been verified efficiently in single-agent reinforcement learning (Zhang et al., 2022). The SAN is biologically reasonable, containing more dynamic neurons, which have shown advantages in dynamic RL tasks with lower energy consumption and better generalization. In this study, we further applied SAN to MARL cooperation scenarios. Our main contributions can be concluded as follows:

1. We are the first to propose the concept of the MoP, which is inspired by the personality theory in psychology, describing a two-step prediction, where the personality estimator (PE) is designed to receive context for estimating the personality of partner under the DPP constraints first, and then behavior prediction is given by the multi-personality network.
2. We incorporate efficient SAN and MoP models to reach multi-scale biological plausibility, where spiking neurons with neural dynamics have been verified efficient in RL-like tasks (Zhang et al., 2022), and we run further to combine neuronal scale dynamics and partner scale cooperations together, to increase the generalization ability of the agent in multi-agent collaboration.
3. The proposed MoP-SAN is then tested in the Overcooked benchmark environment, and the experimental results show a marked better generalization, especially when cooperating with other unseen partners compared to other DNN baselines, which means our proposed algorithm can successfully infer the personality of the unseen partner in the zero-shot collaboration test. We conducted analysis experiments to analyze why the SAN method has better generalization results than DNN baselines.

2. Related works

RL is an essential paradigm in machine learning, which is also suitable for many sequential decision-making tasks. The RL methods have recently achieved good results in many tasks (Silver et al., 2017, 2018; Vinyals et al., 2019). Existing traditional RL methods can be divided into value-based methods (Mnih et al., 2013) and policy-based methods (Schulman et al., 2015). The proposal of the actor-critic method is of milestone significance in RL which combines the advantages of value-based and policy-based methods. Proximal policy optimization (PPO) (Schulman et al., 2017) is one of the most classic methods in this framework, which has achieved compelling performance in many tasks, such as control tasks (Schulman et al., 2017) and StarCraft (Yu et al., 2021).

MARL describes the process of multi-agent learning strategies from scratch to maximize the global rewards in the process of interacting with the environment sequentially or simultaneously. For example, in the two-player cooperative task Overcooked, the ego agent and the partner agent need to cooperate to maximize the team reward from the Overcooked environment. In MARL, cooperative MARL tasks are a very challenging direction. Although there are some studies exploring how to solve challenging problems in cooperative MARL tasks such as credit assignment (Sunehag et al., 2018; Harada et al., 2023), how to design a model which can generalize to unseen partners is still challenging. For multi-agent cooperation, some recent studies (Carroll et al., 2019; Shih et al., 2021, 2022; Strouse et al., 2021; Zhao et al., 2021; Lou et al., 2023) focus on the generalization research of unseen partners. Although traditional self-play methods (Silver et al., 2018) have achieved significant advantages and can often converge to an optimal equilibrium strategy in competitive games, they tend to overfit specific partners for cooperative tasks. Some efforts are put into solving the overfitting through imitation learning (Carroll et al., 2019; Shih et al., 2022) even though it has been reported as challenging in collecting expert data in many real scenarios. For the better generalization of human-AI collaboration, modular methods are proposed, which explicitly separate the convention-dependent representations and rule-dependent representations (Shih et al., 2021). Other studies (Strouse et al., 2021; Zhao et al., 2021) tried to solve the cooperative task of unseen partners by designing various

population pools, which include many carefully designed criteria and agents.

Since brain-inspired SNN has advantages in many aspects (Zhang et al., 2021), many studies have begun to use SNN to solve reinforcement learning problems (Florian, 2007; Frémaux et al., 2013; Patel et al., 2019; Bellec et al., 2020; Tang et al., 2020; Zhang et al., 2022). Our previous study proposed a multi-scale dynamic coding improved the spiking actor network (MDC-SAN) in a single-agent scenario to achieve efficient decision-making (Zhang et al., 2022). Unlike most of these studies that explore SNN methods in single-agent RL tasks, this study wants to apply the SNN method to multi-agent cooperation tasks. In this study, we need to cooperate with different styles of partners in cooperative tasks, so it is vital to construct a model for partner modeling.

ToM (Gallagher and Frith, 2003; Frith and Frith, 2005; Roth et al., 2022; Aru et al., 2023) is a fundamental concept in cognitive psychology, and it allows individuals to predict and explain others' behaviors, communicate effectively, and better engage in cooperative interactions, which is also what we want AI agents to achieve. There are some studies that design ToM models (Tabrez et al., 2020; Wang et al., 2021; Yuan et al., 2022) to solve RL tasks. Through the ToM model, the agent can communicate with other partners more efficiently and learn some conventions for partners. In some studies (Rabinowitz et al., 2018; Roth et al., 2022), the design of the ToM model is to understand the behavior of other agents, which is vital for many RL tasks. While ToM encompasses many aspects, including mental simulation, action prediction, and reasoning, in this context, we will focus on a specific aspect called personality traits in order to enhance the agent model.

3. Method

3.1. The problem setting of 2-player cooperation

We can define this 2-player Markov game as a tuple $(\mathcal{O}, \mathcal{A}, \mathbb{P}, \gamma, \pi, \rho^i, r, m)$, where $\mathcal{O}$ denotes the observation space and $\mathcal{A}$ represents the action space that the ego agent and partner share. We can define $\mathbf{o} = (o^1, o^2)$ including the ego observation and the partner observation. We can denote label $\mathbf{a} = (a^1, a^2)$ as the joint action for all players, including the ego action and the partner action. $\mathbb{P}: \mathcal{O} \times \mathcal{A} \rightarrow \mathcal{O}$ represents the environment transition probability function, and $\gamma \in [0, 1)$ is the discount factor. π is the joint policy, and the policy of ego agent ρ^1 is the spiking policy of the SAN agent for our MoP-SAN, and ρ^2 represents the partner's policy. All agents share the same team reward function $\mathbf{r}(\mathbf{o}, \mathbf{a}): \mathbf{o} \times \mathbf{a} \rightarrow R$. $\tau = (\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, ...)$ denotes the trajectory generated by the joint policy π, and $\tau^2 = (o_0^2, a_0^2, o_1^2, ...)$ is the trajectory of the partner. The MoP model m can model the partner based on the historical trajectory information of the partner and provide actionable guidance for the SAN agent. At each time step, the SAN agent perceives an observation $o_t^1 \in \mathcal{O}$ and receives the guided action $\hat{a}_t^2$ from the MoP model m, taking action $a_t^1 \in \mathcal{A}$ drawn from a spiking policy $\rho^1: \mathcal{O} \times \mathcal{A} \rightarrow [0, 1]$, denoted as $a_t^1 = \rho^1(\cdot|o_t^1, \hat{a}_t^2)$. The policy of the partner can be denoted as $a_t^2 = \rho^2(\cdot|o_t^2)$. The SAN agent and partner enter the next state $\mathbf{o}_{t+1}$ with the probability $\mathbb{P}(\mathbf{o}_{t+1} \mid \mathbf{o}_t, \mathbf{a}_t)$, receiving a numerical reward r_{t+1} from the environment. All agents coordinate together for the maximum cumulative discounted return $\mathbb{E}_{\boldsymbol{\tau} \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{o}_t, \mathbf{a}_t)\right]$.

We assume that there is at least one joint policy through which all agents can attain the maximum cumulative rewards in fully cooperative games. The problem, objective statement, and our approach are formalized in the following sections.

3.2. The algorithmic architecture and pipeline of MoP-SAN

In the last section, the cooperative MARL problem is defined. We present our algorithmic architecture and pipeline for the learning and generalization phases in this section. In this study, we propose a robust framework for multi-agent collaboration. The left side of Figure 1 represents the two phases in our experiment, which will be discussed in the following section. The right side of Figure 1 shows the pipeline of our MoP-SAN in the zero-shot collaboration, and Figure 2 illustrates the detailed structure of our MoP-SAN.

As shown in Figures 1, 2, our proposed framework includes a MoP model and a SAN model as the ego agent under the consideration of biological plausibility and energy efficiency. The MoP as partner mental model can understand the behavior of the partner and model the partner to estimate the personality of partner first and then instruct the action of the SAN agent. The SAN agent can have a better generalization ability of partner heterogeneity (zero-shot collaboration with diverse unseen partners) and cooperate with the unseen partner through the aid of the MoP model m. As shown in Figure 1, we can divide our process into the learning and generalization phases, also called the training and testing process. We introduce a general framework that does not require additional expert-supervised data in the learning phase. In our current model, for simplicity, we assume that the observation encoder is an identity mapping, and the observation from environment is the input to the MoP. In order to self-supervise the training of the MoP model without additional expert data, we directly train MoP as a partner in the learning process for the sake of simplicity.

On the one hand, the MoP model can act as a pool of many diverse agents to facilitate the learning of the SAN agent. On the other hand, the MoP model can also learn various personalities. In the generalization phase, we want to infer better and adapt to the unseen partner with a specific personality, so we need to discover as many base personalities in the personality space as possible during the learning process.

In the generalization phase, parameters in our framework are fixed. As shown in Figure 1, when the SAN agent needs to cooperate with an unseen partner, the personality estimator (PE) determines the partner's personality first according to the historical context information of the unseen partner, and then the multi-personality network infers the current intention and action of the partner. Our goal is to maximize the total reward and entropy based on the historical information of the unseen partner. In the following sections, our descriptions and formulas use the generalization phase as an example to describe our method. The output of our MoP model is the input for the spiking policy of SAN ρ_θ^1 and θ^1 is the parameter for the policy network in SAN. φ and η are the

10.3389/fnins.2023.1219405

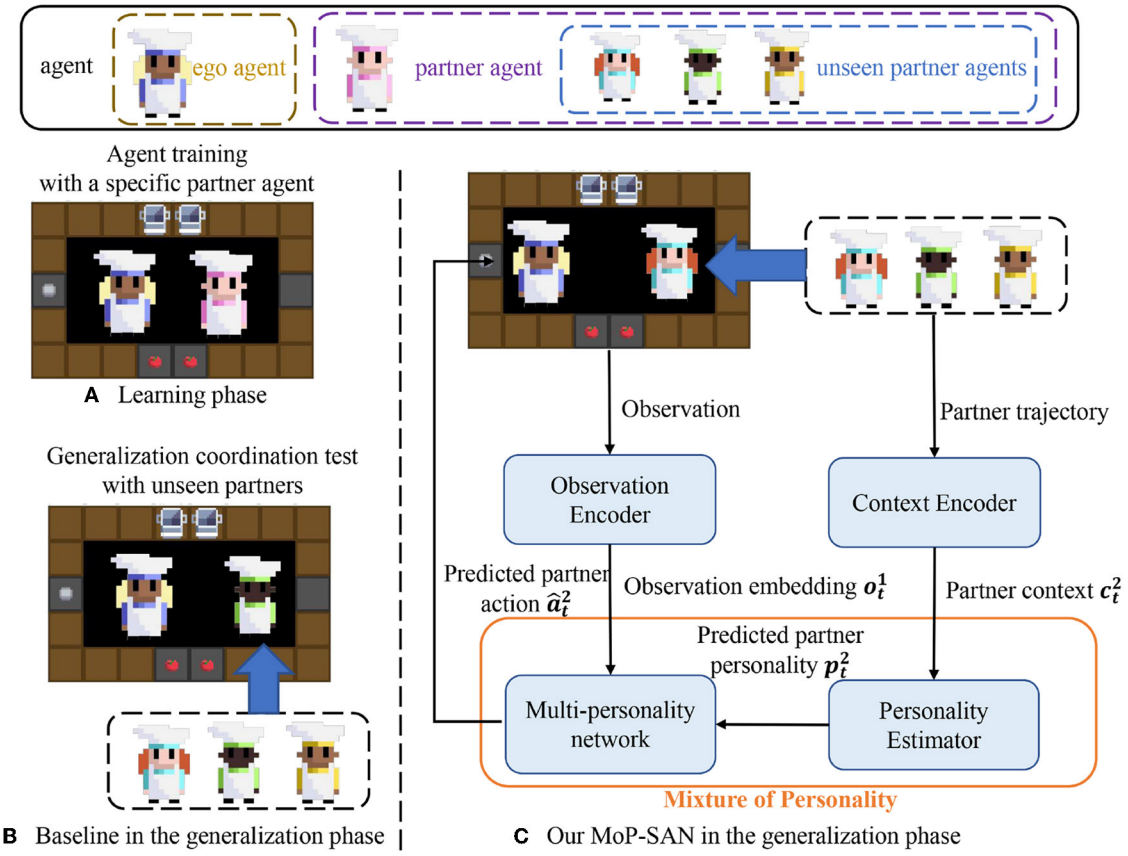

FIGURE 1
Learning and generalization phases of our proposed MoP-SAN and baseline methods. **(A)** Ego agent and the specific partner agent collaborate to complete the Overcooked task in the learning phase. **(B)** In the generalization phase, the ego agent needs to collaborate with some unseen partner agents to test generalization ability (zero-shot collaboration). **(C)** This figure shows our proposed MoP model in the generalization phase. By constructing a MoP model, we can first estimate the partner's personality by the personality estimator and predict the actions of the partner by the multi-personality network according to the personality of the partner. Two agents in the same kitchen in all three graphs represent the cooperative relationship between the two agents to complete this cooking task.

parameter for the MoP model, and the joint policy can be written as follows:

$$\pi\left(\mathbf{a}_t \mid \mathbf{o}_t\right) = \rho_\theta^1\left(a_t^1 \mid o_t^1, \hat{a}_t^2\right) \rho_\theta^2\left(a_t^2 \mid o_t^2\right), \tag{1}$$

where o_t^i is the observation of the i-th player and $\hat{a}_t^2$ denotes the predicted action distribution from our MoP model.

3.3. The SAN model and context encoder

The SAN model in our MoP-SAN refers to a SAN PPO agent, which makes its action based on the guided action of the MoP model to maximize the cooperation reward and entropy. The output action a_t^1 is sampled from the probability distribution over the action space of the spiking policy in the SAN model $\rho_\theta^1\left(a_t^1 \mid o_t^1, \hat{a}_t^2\right)$. The SAN PPO agent includes a spiking actor and critic. The SAN model consists of leaky-integrate-and-fire (LIF) neurons, an abstraction of the Hodgkin-Huxley model. Non-differential membrane potential and refractory period are biologically plausible characteristics of the LIF neuron, which can simulate the neuronal dynamics. We define LIF neurons as follows:

$$\tau \frac{dV(t)}{dt} = -V(t) + I(t), \tag{2}$$

where $V(t)$ represents the dynamic variable of membrane potential for time t and dt is the minimal simulation time slot. $I(t)$ represents the integrated post-synaptic potential and τ is the integrative time period. With input $I(t)$ within a period time of τ when $V(t)$ is bigger than the firing threshold V_{th}, the neuron will be fired and generate a spike, and the membrane potential $V(t)$ will be reset as the reset potential V_{reset}. The neuron will be mostly leaky when $V(t)$ is smaller than the firing threshold. The detailed configuration of SAN is shown in our previous study (Zhang et al., 2022).

The context encoder is the key to our good generalization and adaptation ability. We use the transformer model as our context encoder, and the input of our context encoder is the historical trajectories of the partner in a specific context size as context information. For context information, historical actions and observations have different dimensions. Therefore, we introduce an action MLP network and obs MLP network to convert historical

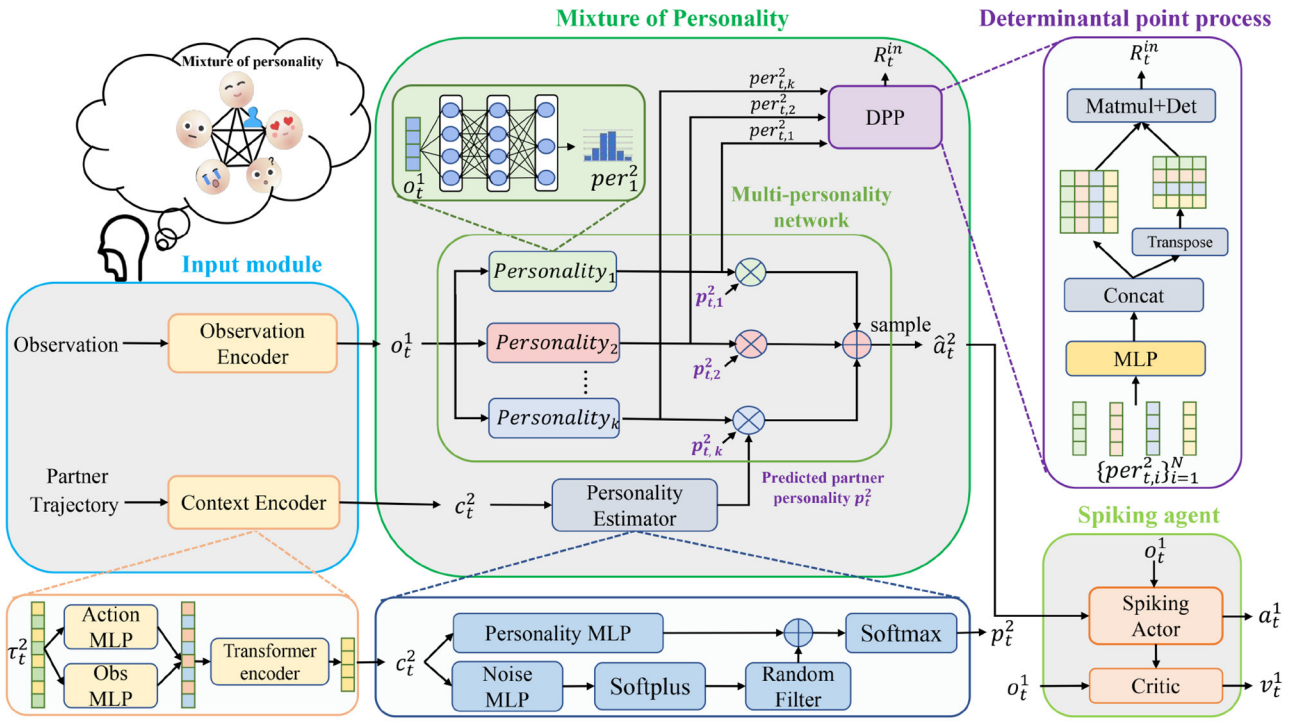

FIGURE 2
Detailed structure of MoP-SAN. MoP-SAN consists of a SAN agent, a MoP model, and an input module that includes a context encoder and an observation encoder. The SAN PPO is used to simulate the ego agent with MoP. The MoP model is used to simulate the theory of mind process of our ego agent modeling the personality of the unseen partners. Our MoP model contains the personality estimator (PE) module, the multi-personality network, and the DPP module.

actions and observations into the same dimension, concatenating them in alternating order according to the order of time t in the trajectory τ, similar to Chen et al. (2021) and Meng et al. (2023).

3.4. The MoP model

The ToM ability of our MoP-SAN is delivered by our MoP model m, which consists of the multi-personality network, the PE module, and the DPP module.

The multi-personality networks include k different personality networks, each consisting of three-layer-MLP that represent a category of base personality strategies with a different policy. The input of our multi-personality network is the observation of the SAN agent, and the output of i-th personality network $per_{t,i}^2$ is a action distribution corresponding to the respective basic personality under the same environmental observation.

The input of the PE module is the partner's context information c_t^2 which is the context embedding from historical trajectories of the partner by context encoder. In contrast to an entirely rational AI agent, the unseen partners are subject to some irrational factors that affect their decisions. Therefore, our PE module consists of a personality multi-layer perceptron (MLP) represented by a trainable weight matrix W_p and a Noise MLP represented by W_{noise}. The output of the Noise MLP is passed through a softplus function and a random filter and then added to the output of the personality MLP. The resulting sum is then passed through a softmax function to obtain an estimated personality profile p_t^2 for an unseen partner. The e represents the PE function and the R denotes a random filter function:

$$e(c_t^2) = \text{Softmax}\left(c_t^2 \cdot W_p + R(c_t^2 \cdot W_{noise})\right), \tag{3}$$

where the output of the MoP model $\hat{a}_t^2$ is sampled from the probability distribution over the action space $m_{\varphi,\eta}\left(\hat{a}_t^2 \mid o_t^1, c_t^2\right)$. The output of the PE module p_t^2 corresponds to the predicted partner personality. η is the parameter of the DPP in MoP and φ is the parameter of the MoP model. The policy of our MoP can be defined as following:

$$m_{\varphi,\eta}\left(\hat{a}_t^2 \mid o_t^1, c_t^2\right) = \sum_{i=1}^{n} p_{t,i}^2 \cdot per_{t,i}^2, \tag{4}$$

where $p_{t,i}^2$ is the i-th coefficient of the output vector of the PE module and $per_{t,i}^2$ represents the output of i-th personality network which is the probability distribution over the action space of the i-th base personality in the current observation. The above equation describes the prediction of our current partner's actions based on the predicted personality of the partner and corresponding actions for a specific personality in the environmental state o_t^1. Instead of a sparsely-activated model that chooses different branches for different tasks, our MoP method integrates the output of all the base personalities rather than selecting a base personality each time. Therefore, the output of the PE module, the predicted personality

of the partner, is not a discrete one-hot vector but a floating-point vector that sums to one.

Our MoP can model partners and infer the personalities of other partners that can help any RL agents to enhance their generalization ability and adaptability so that the agent can be applied to many zero-shot collaboration scenarios.

3.5. The DPP module in the MoP

In this section, we introduce the DPP first and present the DPP in our proposed MoP-SAN. DPP (Kulesza and Taskar, 2012) is an efficient probabilistic model proposed in random matrix theory and has been widely used in many application fields of machine learning (Gong et al., 2014; Parker-Holder et al., 2020; Perez-Nieves et al., 2021), such as recommendation systems (Chen et al., 2018) and video summarization (Gong et al., 2014). The high-performing model DPP can translate complex probability computations into simple determinant calculations and then use the kernel matrix's determinant to calculate the probability of each subgroup. Recent studies, such as Dai et al. (2022) and Yang et al. (2020), have incorporated the DPP model into reinforcement learning (RL) approaches. Dai et al. (2022) utilized DPP models to introduce intrinsic rewards and enhance the exploration of RL methods. Meanwhile, Yang et al. (2020) used DPP to enhance existing RL algorithms by encouraging diversity among agents in RL evolutionary algorithms.

In the learning process, the multi-personality network can be considered to have various personalities. Each personality network can be regarded as a distinct base personality. Measuring the diversity among the multiple base personalities is crucial for constructing a diverse set of base personalities in the personality space. To effectively explore the range of personalities in task space, we integrate a diversity-promoting DPP module to regularize these base personalities in our MoP-SAN. This ensures efficient exploration and optimization of the diverse set of personalities, improving the overall performance of our MoP-SAN.

We can measure the diversity of the personalities and select the subset of diverse personalities through the diversity constraints as an intrinsic reward imposed by the DPP module. Y denotes the set containing many personalities, and y refers to a subset of Y including k personalities that can maximize the diversity. Since these personality networks share the same observation input and the output of a specific personality network $per^2_{t,i}$ is an action distribution, the difference between base personalities can be measured by the action distribution over the action space. We denote the kernel matrix of y as L_y. The determinant value of L_y can represent the diversity of the personality set y. To construct the set y, we need to select k personalities in the personality space for maximizing the determinant value of the kernel matrix of y. The personality set y can be regarded as a set of base personalities that maximizes diversity in the personality space.

$$y^* = \arg\max_y P(Y = y) = \arg\max_y \det\left(L_y\right). \tag{5}$$

Since the matrix L_y is positive semi-definite, there exists matrix B_t at every time step t such that

$$L_y = B_t B_t^T, \tag{6}$$

B_t and the intrinsic reward r_t^{dpp} can be defined as follows, and k is the number of personalities:

$$B_t = \left[\upsilon_\eta\left(per^2_{t,1}\right), \upsilon_\eta\left(per^2_{t,2}\right), \upsilon_\eta\left(per^2_{t,3}\right), \ldots, \upsilon_\eta\left(per^2_{t,k}\right)\right]^T, \tag{7}$$

$$r_t^{\text{dpp}}\left(per^2_{t,1}, per^2_{t,2} \ldots per^2_{t,k}; \eta\right) = \log\det\left(B_t {B_t}^T\right), \tag{8}$$

where υ_η represents the feature vector parameterized by the parameters η.

We endeavor to build some unique personality vectors as base personalities for our multi-personality network, which can combine the entire personality space. Therefore, our MoP model with our proposed DPP module can enable rapid adaptation and generalization to any unseen partners in the collaboration task.

3.6. The SAN learning

The policy parameters of the SAN agent θ^1 and the MoP model parameter (φ, η) are iteratively optimized in our method. The overall optimization objective is to maximize the cumulative discounted return, which depends on the MoP model $m_{\varphi,\eta}\left(a_t^2 \mid o_t^2, c_t^2\right)$ and the spiking policy of the SAN agent $\rho_\theta^1\left(a_t^1 \mid o_t^1, a_t^2\right)$:

$$\theta^{1*}, \varphi^*, \eta^* = \max_{\theta,\varphi,\eta} \sum_{t=0}^{\infty} \mathbb{E}_{a_t^1, a_t^2}\left[\gamma^t\left(r\left(\mathbf{o}_t, \mathbf{a}_t\right) + \alpha\overline{\mathcal{H}}\left(\pi\left(\mathbf{a}_t \mid \mathbf{o}_t\right)\right)\right)\right]. \tag{9}$$

The goal of the SAN agent is to maximize the extrinsic reward r_t^{ex} by collaborating with partners. We can calculate the gradient of the SAN as follows:

$$\begin{aligned} \nabla_\theta J\left(\rho_\theta^1\right) = \mathbb{E}_{a_t^1, a_t^2}\Big[&\nabla_\theta \log(\rho_\theta^1(a_t^1|o_t^1, a_t^2))\big(G^{ex}(\mathbf{o}_t, \mathbf{a}_t) \\ &- b_1\left(o_t^1, \mathbf{a}_t\right) - \alpha\log(\rho_\theta^1(a_t^1 \mid o_t^1, a_t^2))\big)\Big] \\ a_t^1 \sim \rho_\theta^1\left(a_t^1 \mid o_t^1, a_t^2\right), &a_t^2 \sim m_{\varphi,\eta}\left(a_t^2 \mid o_t^2, c_t^2\right) \end{aligned}, \tag{10}$$

where the b_1 is the baseline function and $G^{\text{ex}}\left(\mathbf{o}_t, \mathbf{a}_t\right)$ denotes the discounted extrinsic returns for SAN. In the study, we used the game score as the extrinsic reward r_t^{ex}. The above equation describes the optimization process for the ego SAN agent in our MoP-SAN method similar to the PPO optimization (Schulman et al., 2017) in RL. We can estimate the baseline function b_1 by the expected return of all possible actions, as shown in follows:

$$b_1\left(o_t^1, a_t^1\right) = \sum_{a_t^1 \in \mathcal{A}} \rho_\theta^1\left(a_t^1 \mid o_t^1, a_t^2\right) G^{ex}\left(\mathbf{o}_t, \mathbf{a}_t\right). \tag{11}$$

3.7. The MoP learning

We introduced the DPP constraint into our study, similar to a recent study (Dai et al., 2022), by treating the DPP diversity measurement as the intrinsic reward. We adopted a bi-level optimization framework (Dai et al., 2022) for the MoP model and its DPP module to maximize the intrinsic reward and extrinsic reward.

Our objective can be defined as follows:

$$\max_{\eta} J^{\mathrm{ex}}\left(\varphi', \eta\right) \text{ s.t. } \varphi' = \underset{\varphi}{argmax} J^{\mathrm{mix}}(\varphi, \eta), \tag{12}$$

for this optimization problem, we can treat it as a Stackelberg game. We use the DPP reward as the intrinsic reward. The mixture rewards are the sum of intrinsic and extrinsic rewards. The mixture reward can be written as follows:

$$r_t^{\mathrm{mix}} = r_t^{\mathrm{ex}} + \beta r_t^{\mathrm{dpp}}\left(a_1, a_2 \ldots a_k; \eta\right), \tag{13}$$

where β is the weight coefficient of the intrinsic reward. r_t^{ex} is the standard reward from the environment where the SAN agent makes actions a_t^1, and MoP makes a_t^2 in the environmental state s_t at the time step t, and r_t^{dpp} is the DPP constraint diversity reward for the partner. The gradient $\nabla_\varphi J^{mix}$ can be calculated as follows:

$$\begin{aligned}\nabla_\varphi J^{\mathrm{mix}} = \alpha \cdot \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 \mid o_t^2, c_t^2\right)\Big(G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right) \\ - b_2\left(o_t^2, \mathbf{a_t}\right) - \alpha \log\left(m_{\varphi,\eta}\left(a_t^2 \mid o_t^2, c_t^2\right)\right)\Big)\end{aligned}, \tag{14}$$

where $G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right)$ denotes the discounted mixture returns for our MoP-SAN. The gradient $\nabla_\eta J^{\mathrm{ex}}$ can be calculated by using the chain rule:

$$\nabla_\eta J^{\mathrm{ex}} = \nabla_{\varphi'} J^{\mathrm{ex}} \nabla_\eta \varphi', \tag{15}$$

with

$$\begin{aligned}\nabla_\eta \varphi' &= \nabla_\eta \alpha G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right) \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right) \\ &= \alpha\beta \sum_{l=0}^{\infty} \gamma^l \nabla_\eta R_{\eta,t+l}^{\mathrm{d}} \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right).\end{aligned} \tag{16}$$

We can use importance sampling to improve the sample efficiency of the algorithm:

$$\nabla_{\varphi'} J^{\mathrm{ex}} = \nabla_{\varphi'}\left(\frac{m_{\varphi',\eta}\left(a_t^2 | o_t^2, c_t^2\right)}{m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right)}\right) G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right), \tag{17}$$

$$\begin{aligned}\nabla_\eta J^{\mathrm{ex}} &= \nabla_{\varphi'} J^{\mathrm{ex}} \nabla_\eta \varphi' \\ &= \nabla_{\varphi'}\left(\frac{m_{\varphi',\eta}\left(a_t^2 | o_t^2, c_t^2\right)}{m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right)}\right) G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right) \alpha\beta \cdot \\ &\sum_{l=0}^{\infty} \gamma^l \nabla_\eta R_{\eta,t+l}^{\mathrm{dpp}} \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right)\end{aligned}. \tag{18}$$

Hence, the iterative learning of policy parameters in the SAN and MoP model finally converges the whole system to support next-step MARL tasks.

4. Experimental results

4.1. Environmental settings

Our experimental environment is Overcooked (Carroll et al., 2019), a primary human-AI zero-shot collaboration benchmark. Similar to previous studies (Carroll et al., 2019; Shih et al., 2021, 2022), we have conducted experiments on the "simple" map based on PantheonRL (Sarkar et al., 2022), a pytorch framework for human-AI collaboration. In this environment, two players cooperate to complete the cooking task, i.e., making as many onion soups as possible for winning a higher reward in a limited time. The players can choose one of six actions and execute simultaneously, including up, down, left, and right, empty operation, or interaction.

It is necessary to follow a specific order when making onion soup. The player must put three onions in the pot and cook them for 20 steps. Then player pours the onion soup from the pot onto the plate and serves the dish to the designated position. After this process, the player can get certain rewards (20). A player can not complete this task alone on the challenging task. Only through good cooperation can the players achieve high scores, which requires the ability to infer the personality of the partner first and predict the actions of the partner.

4.2. Configurations of our baselines and our MoP-SAN

There are several baseline methods. One method is the standard DNN PPO baseline (Schulman et al., 2017), an important MARL method with excellent performance in many scenarios. In this method, both ego and partner agent are homogeneous PPO agents, and this way is also called self-play (Silver et al., 2018) in RL.

Another important baseline is the SAN PPO baseline. In this study, we choose SAN as our baseline for three main reasons. The first reason is that SAN is the ego agent in our MoP-SAN method, and our MoP model serves as a ToM model to provide partner action predictions for SAN. Other reasons include the higher generalization performance for one-shot learning and the improvement of energy efficiency. Since the ego agent in our method is also the SAN PPO, we refer to the SAN PPO baseline as the SAN baseline in the following experimental description. It is worth mentioning that we first introduce the SAN version of PPO into the multi-agent cooperation task Overcooked. For the SAN baseline, in our cooperation environment, the ego agent is the SAN PPO, and the partner is the standard PPO.

The experimental details of our setting are shown in Figure 3. As shown in Figures 1, 3, the SAN agent and MoP in one pair have the same name and are trained together by iterative optimization in the learning phase for our MoP-SAN. For example, our SAN A as the ego agent and MoP A as the partner will cooperate in the learning phase for a good score. In the generalization phase, SAN and MoP with the same name will be combined into MoP-SAN as the ego agent. We will evaluate the generalization of our proposed

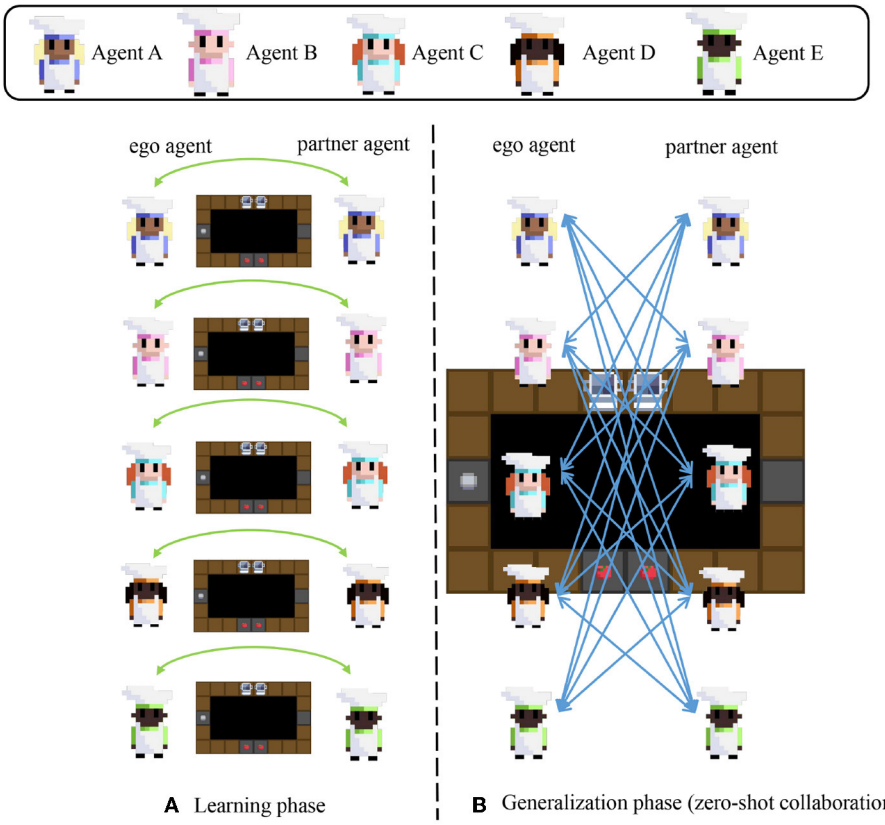

FIGURE 3
Experimental setting detail for our MoP-SAN in learning and generalization phases (zero-shot collaboration). Agent A-E corresponds to the agent with different seeds whose name is A-E. **(A)** In the learning phase, the ego agent and specific partner agent in a pair collaborate for this task and are trained by iterative optimization. The ego agent and partner agent in a pair have the same name. There are five agent pairs in the learning phase: (A, A), (B, B), (C, C), (D, D), and (E, E). **(B)** In the generalization phase, the ego agent needs to collaborate with all unseen partner agents in a zero-shot manner. For example, the ego agent A will cooperate with another unseen partner agent with a different name (B, C, D, or E) for the zero-shot collaboration test.

MoP-SAN model by cooperating with different unseen partners, which means the ego and partner agent in one pair have different names.

Our training experiment is run for half a million steps, and the generalization experiment (zero-shot collaboration) is conducted for several games to take the average score during the generalization phase in all our experiments. The personality number is 12, and the context size is 5. For the context encoder in our MoP-SAN, if the length of historical trajectories of the partner is less than the context size, we will pad 0. We use a single-layer transformer with two heads as a context encoder whose inner dimension is 256 and the dimension for q,k,v is 64. For the part of padding 0, we mask it in the transformer. Our MoP-SAN model uses an actor-critic framework, and the actor is based on SAN, similar to a previous study (Zhang et al., 2022). The actor network is (64, tanh, 64, tanh, 6); the critic network is (64, tanh, 64, tanh, 1). We sample action from categorical distribution for all methods. In these methods, we use the Adam optimizer, and the learning rate is 0.0003. The reward discount factor is $\gamma = 0.99$, and the batch size is 64. The weight coefficient of the intrinsic reward β is 0.5, and the maximum length of the replay buffer is 2048. We use gradient clipping to prevent exploding and vanishing gradients.

4.3. Stronger generalization ability of MoP-SAN

Figure 4 is a histogram representing the generalization and learning scores obtained by three methods in the Overcooked task. The line chart in the histogram shows the trend of the average score for the different methods. The red dot indicates the average score of all corresponding agents, and the shaded area represents the standard deviation of the corresponding results for the three methods.

The average score for the method in the left diagram is the average score of all generalization tests with unseen partners. As shown in Figure 3, the average score for our MoP-SAN method in A is 142, which means that the average for four unseen tests (A-B, A-C, A-D, and A-E) is 142. The average score for our method is 142.25 means that the average for twenty unseen tests (A-B, A-C, A-D, A-E, B-A, B-C, B-D, B-E, C-A. . .) is 142.25. Figure 5 shows the detailed score for all generalization tests with unseen partners. The detailed score in the learning and generalization phase for each pair can be found in the Supplementary material.

Figure 4 indicates that our proposed MoP-SAN model outperforms all baselines for unseen partners during the zero-shot collaboration, showing a more robust and stable ability for

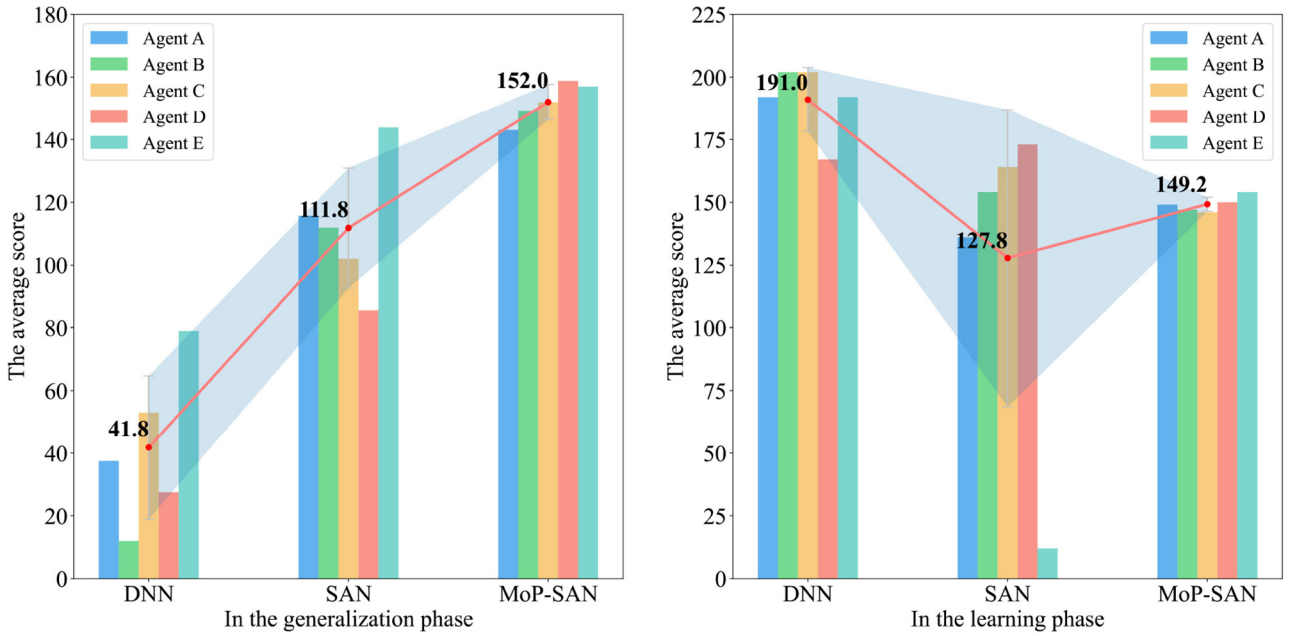

FIGURE 4
Score comparison between the baseline and MoP-SAN models in the generalization and learning phases. The **(left)** figure shows that our MoP-SAN outperforms other baselines in terms of generalization performance, with a strong generalization ability to complete cooperative tasks with unseen partners. The **(right)** figure shows that our MoP-SAN improves the poor performance and large variance of SAN in the learning phase. Agents A-E denote different agents with different random seeds.

cooperation. What needs to be further emphasized is that our MoP-SAN method not only significantly outperforms the SAN baseline but also the DNN baseline in the generalization test, which strongly demonstrates the powerful generalization ability for partner heterogeneity of our method in zero-shot collaboration.

The average score in the learning phase can be found in the right diagram of the Figure 4. Although our MoP-SAN method primarily focuses on zero-shot generalization test without any prior knowledge of partners, the scores during the learning phase can still reflect the collaborative performance with the specific partner. Our MoP-SAN has better learning scores and minor variance compared to the SAN baseline in the learning phase.

4.4. Significantly better zero-shot collaborative performance of MoP-SAN

Our experimental results in the zero-shot collaboration test reflect the generalization ability of partner heterogeneity for different methods. Figure 5 is the color temperature map showing the specific experimental data in the generalization test for all three methods. The color temperature maps in Figure 5 correspond to the DNN baseline, the SAN baseline, and our MoP-SAN model, respectively. The row represents the ego agent, and the column represents the partner. For example, the score in the first row, the third column for our MoP-SAN represents the zero-shot collaboration score between MoP-SAN A and unseen partner C. The scores on the diagonal represent the scores achieved by the corresponding pairs during the learning phase, which are not included in the zero-shot collaboration score data of the generalization phase. We can see that the more obvious the color difference is, the more significant the variance of this method.

As shown in Figure 5, our multi-scale biological plausibility MoP-SAN achieved significantly better scores and smaller variance than the other baselines for most pairs in the zero-shot generalization test with low energy consumption, achieving good generalization results with unseen partners of different styles. As shown in Figure 6, although DNN achieves high scores in some generalization test experiments, its variance is large, and the average score is low. Moreover, the SAN baseline has a better average score and smaller variance than the DNN baseline. These results demonstrate that our MoP model can complete partner modeling and help the SAN agent have a higher collaborative score with a better generalization ability.

The question of why SAN can achieve better generalization results than DNN has caught our attention. In order to further verify whether the poor generalization test performance of DNN was due to overfitting, we conducted a series of analysis experiments on DNN. We saved the checkpoints of DNN's learning process from underfitting to "overfitting" and performed unseen partner generalization tests. As shown in Figure 6, these results indicate that as the number of training steps increases, the generalization performance of DNN gradually improves. We have discovered a similar pattern in these test results and named it the DNN type.

Similarly, in the generalization test results of SAN, we also discovered a similar pattern which we named the SAN type. As shown in Figure 6, compared to the DNN type, the SAN type exhibits stronger generalization and cooperation abilities in

10.3389/fnins.2023.1219405

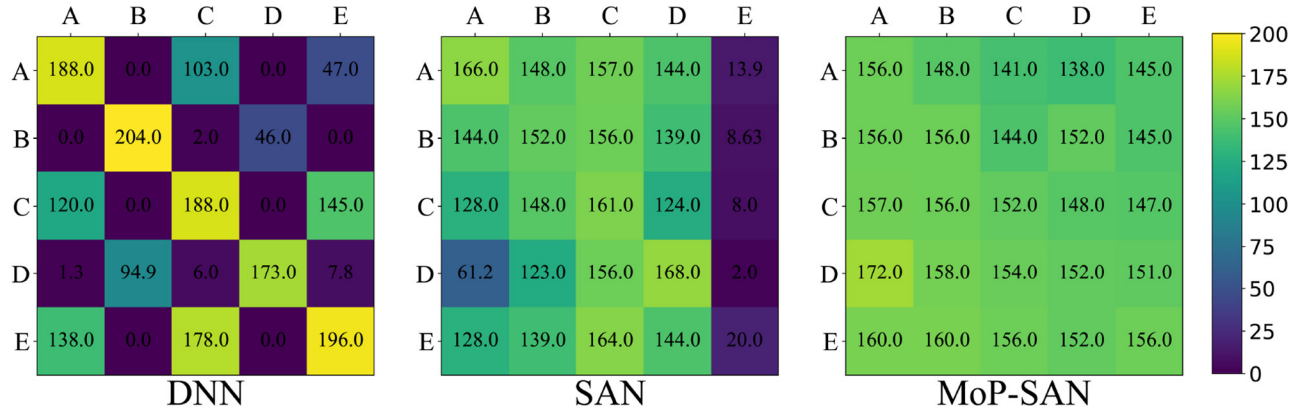

FIGURE 5
Color temperature diagram shows the detailed generalization score for the baseline methods and our MoP-SAN. The difference in colors demonstrates the difference in scores. Compared with the DNN and SAN baseline, our proposed MoP-SAN has more satisfactory results for a better score and smaller variance.

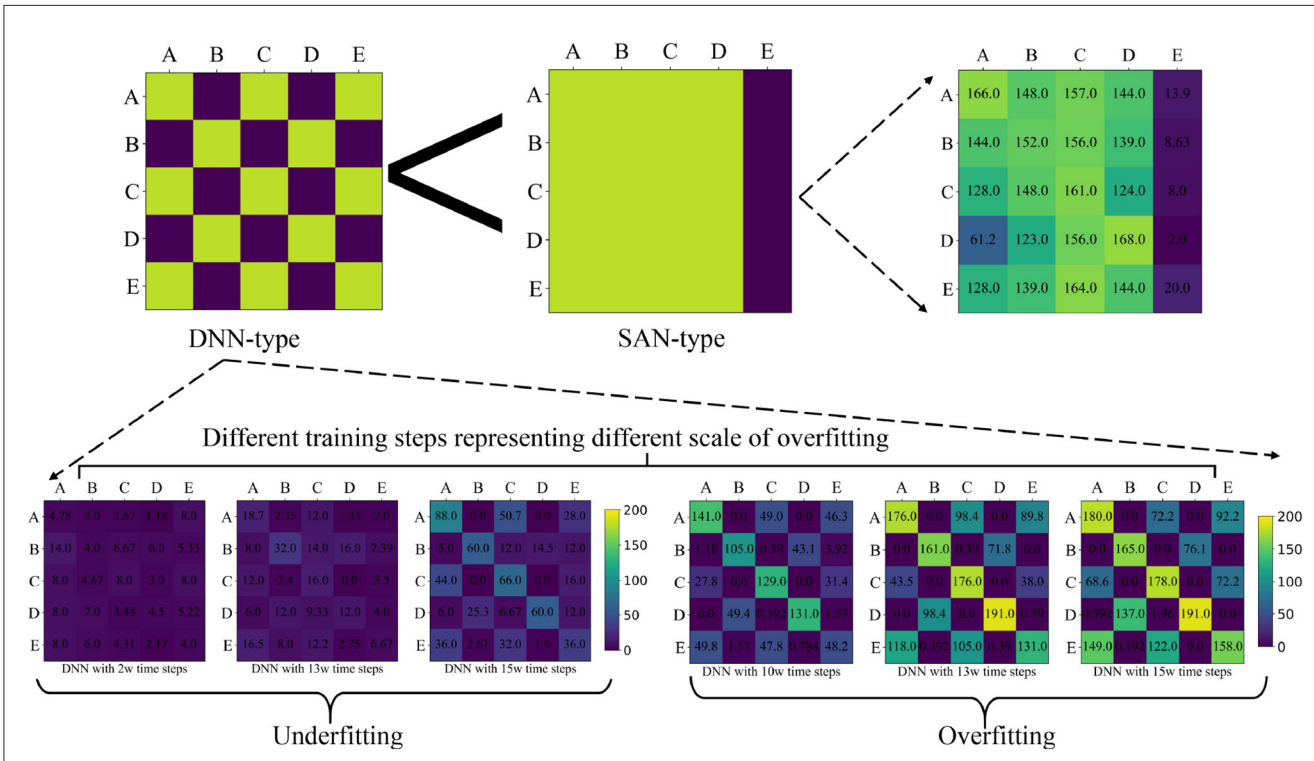

FIGURE 6
Diagram depicts the detailed generalization analysis experiment of DNN and SAN, showing the generalization test results of the DNN under different training steps, which represent different scales of overfitting. As the number of training steps increases, the generalization performance of DNN gradually improves. The generalization test results for DNN exhibit a similar pattern of DNN-type, while the results for SAN also exhibit a similar pattern of SAN-type. By comparing these two patterns, we can see that SAN has better generalization ability and robustness.

unseen partner generalization scenarios. These results represent that "overfitting" was not the main cause of the poor generalization test performance of DNN. We believe that the reason why DNN performs worse than SAN in the generalization test with unseen partners is that SAN has better noise resistance and robustness. In cooperative reinforcement learning, the generalization test with unseen partners can be regarded as a noise perturbation test, and therefore, SAN performs better than DNN in our generalization experiment.

4.5. Larger personality size contributes better cooperative performance

Furthermore, we conduct some ablation experiments to confirm the effectiveness of different modules and parameters in our MoP-SAN. The experimental results in Table 1 show that as the number of personalities increases, the learning ability of our MoP-SAN model gradually improves and the variance gradually

gets smaller. These results also show that diverse personalities play an essential role in the multi-agent cooperation task.

From Table 1, we can see that some pairs have very poor cooperation scores when the number of base personalities is small. This may be because these base personalities can not be combined to express all the dimensions of the personality of the partners. As the number of base personalities increases, the expression ability of the existing base personalities for personality of the current partner grows, resulting in better performance.

The personality theory in cognitive psychology suggests that breaking down personality into finer-grained traits is an excellent way to improve predicting and explaining human behavior. Our experimental results further validate this point. By using a larger personality number, we obtain more precise personality delineation, which can better predict the personality of the partner and cooperate more efficiently with partners to achieve higher scores.

TABLE 1 Mean score of different number of personalities in our method.

Agents	A	B	C	D	E	Avg
Ours w/personality 6	0.2	0.4	0.4	0	1.6	0.52 (±0.63)
Ours w/personality 8	1.8	123	0	0.4	0	25.04 (±54.77)
Ours w/personality 10	7.6	151	1.6	**157**	114	86.24 (±76.35)
Ours w/personality 12	**149**	**154**	**150**	146	**146**	**149 (±3.32)**

Bold values indicate the setting which can produce the best results, i.e., the maximum value in that column, facilitating comparisons between the results.

TABLE 2 Mean score of different number of context size in our method.

Agents	A	B	C	D	E	Avg
Ours w/context 0	4.4	2.8	2.6	1	1.2	2.4 (±1.38)
Ours w/context 1	11.6	0.2	1	17.2	85.2	23.04 (±35.48)
Ours w/context 3	136	148	143	140	140	141.4 (±4.45)
Ours w/context 5	**149**	**154**	**150**	**146**	**146**	**149 (±3.32)**

Bold values indicate the setting which can produce the best results, i.e., the maximum value in that column, facilitating comparisons between the results.

4.6. Richer context information contributes better personality prediction

Table 2 indicates that as the context information of the partner increases, the score of our MoP-SAN in the learning phase gets better and better, which shows that partner information is crucial for our MoP-SAN model in the cooperation task. The result is the worst when there is no partner information at all. This is because partner information serves as input for the PE module to predict the personality of partner. Without such information, the personality prediction is random, leading to inefficient collaboration between ego and partner agents when completing tasks such as making onion soup. Limited partner information may make the personality prediction inaccurate, which is detrimental to the collaboration score.

These results in Table 2 also indicate that the existence of partner context information is the key to our ability to solve this task. We find that the existence of partner information achieves better results in the learning phase and gets better generalization results in the zero-shot collaboration generalization experiment.

4.7. Personality diversity controlled by DPP

The results in the ablation experiment of DPP demonstrate the effectiveness of the DPP module, which can achieve better

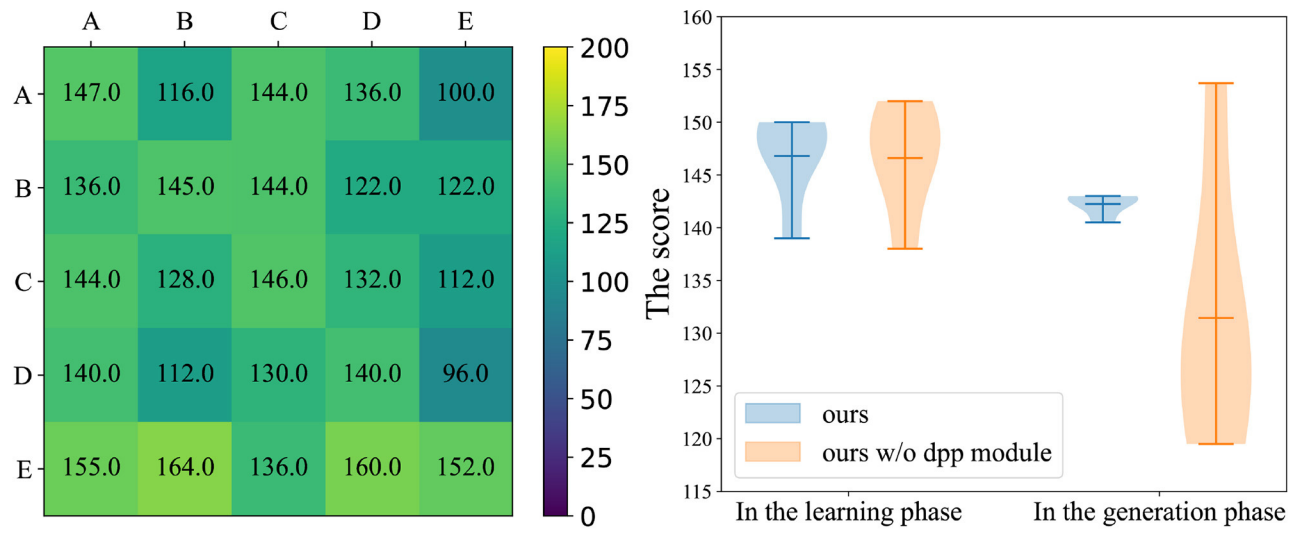

FIGURE 7
Left color temperature diagram demonstrates the detailed generalization scores for our method w/o the DPP module. Right violin plot demonstrates the visual comparison of the scores in the ablation experiments on the DPP module, where violin plots are presented for our method w/o the DPP module and our method for the learning and generalization phases.

results in the generalization experiments. We further analyze the results of the ablation experiment through the color temperature map and violin plot in Figure 7. We show the maximum, minimum, and average lines in the violin plot, and the shade means the data distribution whose size represents the variance of the corresponding method. As shown in the right violin plot of Figure 7, our method is much better than our method w/o DPP at the generalization test, and our MoP-SAN has a smaller variance than our MoP-SAN w/o DPP. The color temperature plot of our MoP-SAN is shown in Figure 5 as the third plot c. The comparison between the left color diagram in Figure 7 with plot c in Figure 5 indicates that our MoP-SAN model has better generalization performance and minor variance owing to the DPP module.

This result indicates that with the same size of personality number, the addition of DPP can constrain the base personalities in MoP, which allows these base personalities to cover as much personality space as possible. This complete coverage leads to a more robust PE module that can more accurately predict the personality of unseen partner, achieving in better scores.

5. Conclusion

In this study, we focus on strengthening the conventional actor network by incorporating multi-scale biological inspirations, including the local scale neuronal dynamics with spike encoding and global scale personality theory with the spirit of the theory of mind. Our proposed mixture of the personality improved the spiking actor-network (MoP-SAN) algorithm can remarkably improve the generalization and adaptability in the MARL cooperation scenarios under a surprisingly low energy consumption.

Our MoP-SAN is then verified by experiments, which shows that the two-step process in personality theory is very crucial for predicting the unseen partner's actions. The MoP improved SAN shows a more satisfactory learning ability and generalization performance compared with SAN and DNN baseline. To the best of our knowledge, we are the first to apply SAN and MoP in the MARL cooperation task. This integrative success has given us more confidence about borrowing more inspirations from neuroscience and cognitive psychology in future for designing new-generation MARL algorithms.

Although the biologically plausible MoP-SAN approach can improve collaboration efficiency and scores in two-player cooperative tasks, our MoP-SAN method can not achieve significant results when cooperating with seen partners, and the complex module design resulted in some computational overhead. It is worth exploring how to apply biological and cognitive inspirations to enhance collaboration efficiency among three or more players. Additionally, it is also worth investigating how to collaborate better with non-rational players.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

BX, JS, TZ, and XL gave the idea. XL and ZN made the experiments and the result analyses. XL, JR, and LM were involved in problem definition. All authors wrote the study together and approved the submitted version.

Funding

This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA27010404 and XDB32070100), the Open Fund/Postdoctoral Fund of the Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences (Grant No.: CoDeCS-OPF-XDA27040807), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX), and the Youth Innovation Promotion Association CAS.

Acknowledgments

The authors would like to thank Yali Du, Dengpeng Xing, Zheng Tian, and Duzhen Zhang for their previous assistance with the valuable discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2023.1219405/full#supplementary-material

References

Anglim, J., and Horwood, S. (2021). Effect of the COVID-19 pandemic and big five personality on subjective and psychological well-being. *Soc. Psychol. Pers. Sci.* 12, 1527–1537. doi: 10.1177/1948550620983047

Aru, J., Labash, A., Corcoll, O., and Vicente, R. (2023). Mind the gap: challenges of deep learning approaches to theory of mind. *Artif. Intell. Rev.* 1–16. doi: 10.1007/s10462-023-10401-x

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., et al. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nat. Commun.* 11, 1–15. doi: 10.1038/s41467-020-17236-y

Boyd, R., and Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 3281–3288. doi: 10.1098/rstb.2009.0134

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia. S., Abbeel, P., et al. (2019). "On the utility of learning about humans for Human-AI coordination," in *Advances in Neural Information Processing Systems*, Vol. 32, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. Alche-Buc, E. Fox, and R. Garnett (Curran Associates, Inc). Available online at: https://proceedings.neurips.cc/paper_files/paper/2019/file/f5b1b89d98b7286673128a5fb112cb9a-Paper.pdf

Cattell, H. E., and Mead, A. D. (2008). The sixteen personality factor questionnaire (16PF). *SAGE Handb. Pers. Theory Assess.* 2, 135–159. doi: 10.4135/9781849200479.n7

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., et al. (2021). "Decision transformer: Reinforcement learning via sequence modeling," in *Advances in Neural Information Processing Systems*, Vol. 34, eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Curran Associates), 15084-15097. Available online at: https://proceedings.neurips.cc/paper_files/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf

Chen, L., Zhang, G., and Zhou, E. (2018). "Fast greedy MAP inference for determinantal point process to improve recommendation diversity," in *Advances in Neural Information Processing Systems*, Vol. 31, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Curran Associates, Inc).

Dai, T., Du, Y., Fang, M., and Bharath, A. A. (2022). Diversity-augmented intrinsic motivation for deep reinforcement learning. *Neurocomputing* 468, 396–406. doi: 10.1016/j.neucom.2021.10.040

De Raad, B. (2000). *The Big Five Personality Factors: The Psycholexical Approach to Personality*. Hogrefe & Huber Publishers.

Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Comput.* 19, 1468–1502. doi: 10.1162/neco.2007.19.6.1468

Frémaux, N., Sprekeler, H., and Gerstner, W. (2013). Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS Comput. Biol.* 9, e1003024. doi: 10.1371/journal.pcbi.1003024

Frith, C., and Frith, U. (2005). Theory of mind. *Curr. Biol.* 15, R644–R645. doi: 10.1016/j.cub.2005.08.041

Gallagher, H. L., and Frith, C. D. (2003). Functional imaging of "theory of mind". *Trends Cogn. Sci.* 7, 77–83. doi: 10.1016/S1364-6613(02)00025-6

Gong, B., Chao, W. -L., Grauman, K., and Sha, F. (2014). "Diverse sequential subset selection for supervised video summarization," in *Advances in Neural Information Processing Systems*, Vol. 27, eds Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Curran Associates, Inc). Available online at: https://proceedings.neurips.cc/paper_files/paper/2014/file/0eec27c419d0fe24e53c90338cdc8bc6-Paper.pdf

Harada, T., Matsuoka, J., and Hattori, K. (2023). Behavior analysis of emergent rule discovery for cooperative automated driving using deep reinforcement learning. *Artif. Life Robot.* 28, 31–42.

Kulesza, A., and Taskar, B. (2012). Determinantal point processes for machine learning. *Found. Trends Mach. Learn.* 5, 123–286. doi: 10.1561/2200000044

Lou, X., Guo, J., Zhang, J., Wang, J., Huang, K., and Du, Y. (2023). Pecan: leveraging policy ensemble for context-aware zero-shot human-ai coordination. *arXiv preprint arXiv:2301.06387*.

McCrae, R. R., and Costa, P. T. Jr. (2008). "The five-factor theory of personality," in *Handbook of Personality: Theory and Research*, eds O. P. John, R. W. Robins, and L. A. Pervin (The Guilford Press), 159–181.

Meng, L., Wen, M., Le, C., Li, X., Xing, D., Zhang, W., et al. (2023). Offline pre-trained multi-agent decision transformer. *Mach. Intell. Res.* 20, 233–248. doi: 10.1007/s11633-022-1383-7

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., et al. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

O'Connor, M. C., and Paunonen, S. V. (2007). Big five personality predictors of post-secondary academic performance. *Pers. Individ. Diff.* 43, 971–990. doi: 10.1016/j.paid.2007.03.017

Parker-Holder, J., Pacchiano, A., Choromanski, K. M., and Roberts, S. J. (2020). "Effective diversity in population based reinforcement learning," in *Advances in Neural Information Processing Systems*, Vol. 33, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates), 18050–18062. Available online at: https://proceedings.neurips.cc/paper_files/paper/2020/file/d1dc3a8270a6f9394f88847d7f0050cf-Paper.pdf

Patel, D., Hazan, H., Saunders, D. J., Siegelmann, H. T., and Kozma, R. (2019). Improved robustness of reinforcement learning policies upon conversion to spiking neuronal network platforms applied to atari breakout game. *Neural Netw.* 120, 108–115. doi: 10.1016/j.neunet.2019.08.009

Perez-Nieves, N., Yang, Y., Slumbers, O., Mguni, D. H., Wen, Y., and Wang, J. (2021). "Modelling behavioural diversity for learning in open-ended games," in *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, eds M. Meila and T. Zhang (PMLR), 8514-8524. Available online at: http://proceedings.mlr.press/v139/perez-nieves21a/perez-nieves21a.pdf

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., and Botvinick, M. (2018). "Machine theory of mind," in *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, eds J. Dy and A. Krause (PMLR), 4218–4227. Available online at: http://proceedings.mlr.press/v80/rabinowitz18a/rabinowitz18a.pdf

Rand, D. G., and Nowak, M. A. (2013). Human cooperation. *Trends Cogn. Sci.* 17, 413–425. doi: 10.1016/j.tics.2013.06.003

Roth, M., Marsella, S., and Barsalou, L. (2022). "Cutting corners in theory of mind," in *Proceedings of the AAAI 2022 Fall Symposium Series on Thinking Fast and Slow and Other Cognitive Theories in AI* (Arlington, TX: Westin Arlington Gateway). Available online at: https://ceur-ws.org/Vol-3332/paper11.pdf

Ruan, J., Du, Y., Xiong, X., Xing, D., Li, X., Meng, L., et al. (2022). "GCS: Graph-based coordination strategy for multi-agent reinforcement learning," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS (ACM Press), 1128–1136.

Ryckman, R. M. (2012). *Theories of Personality*. Cengage Learning.

Sarkar, B., Talati, A., Shih, A., and Sadigh, D. (2022). "PantheonRL: A MARL library for dynamic training interactions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36 (AAAI Press), 13221–13223. doi: 10.1609/aaai.v36i11.21734

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). "Trust region policy optimization," in *International Conference on Machine Learning* (PMLR), 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schultz, D. P., and Schultz, S. E. (2016). *Theories of Personality*. Cengage Learning.

Shih, A., Ermon, S., and Sadigh, D. (2022). "Conditional imitation learning for multi-agent games," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE), 166–175. doi: 10.1109/HRI53351.2022.9889671

Shih, A., and Swahney, A. (2021). "On the critical role of conventions in adaptive human-Ai collaboration," in *International Conference on Representation Learning*.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362, 1140–1144. doi: 10.1126/science.aar6404

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270

Strouse, D. J., McKee, K., Botvinick, M., Hughes, E., and Everett, R. (2021). "Collaborating with humans without human data," in *Advances in Neural Information Processing Systems*, Vol. 34, eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Curran Associates), 14502-14515. Available online at: https://proceedings.neurips.cc/paper_files/paper/2021/file/797134c3e42371bb4979a462eb2f042a-Paper.pdf

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., et al. (2018). "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proceedings of the 17th International Conference on Autonomous Agents and Multi Agent Systems*. 2085–2087.

Tabrez, A., Luebbers, M. B., and Hayes, B. (2020). A survey of mental modeling techniques in human–robot teaming. *Curr. Robot. Rep.* 1, 259–267. doi: 10.1007/s43154-020-00019-0

Tang, G., Kumar, N., and Michmizos, K. P. (2020). "Reinforcement co-learning of deep and spiking neural networks for energy-efficient mapless navigation with neuromorphic hardware," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 6090–6097. doi: 10.1109/IROS45743.2020.9340948

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, Vol. 30, eds I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc). Available online at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., et al. (2019). Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nature* 575, 350–354. doi: 10.1038/s41586-019-1724-z

Wang, Y., Xu, J., and Wang, Y. (2021). "ToM2C: Target-oriented multi-agent communication and cooperation with theory of mind," in *International Conference on Learning Representations*.

Yang, R., Xu, H., Wu, Y., and Wang, X. (2020). "Multi-task reinforcement learning with soft modularization," in *Advances in Neural Information Processing Systems*, Vol. 33, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates), 4767–4777. Available online at: https://proceedings.neurips.cc/paper_files/paper/2020/file/32cfdce9631d8c7906e8e9d6e68b514b-Paper.pdf

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., et al. (2021). "The surprising effectiveness of PPO in cooperative multi-agent games," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., et al. (2022). *In situ* bidirectional human-robot value alignment. *Sci. Robot.* 7, eabm4183. doi: 10.1126/scirobotics.abm4183

Zhang, D., Zhang, T., Jia, S., and Xu, B. (2022). "Multi-scale dynamic coding improved spiking actor network for reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36 (AAAI Press), 59–67. doi: 10.1609/aaai.v36i1.19879

Zhang, M., Wang, J., Wu, J., Belatreche, A., Amornpaisannon, B., Zhang, Z., et al. (2021). Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 1947–1958.

Zhao, R., Song, J., Haifeng, H., Gao, Y., Wu, Y., Sun, Z., et al. (2021). Maximum entropy population based training for zero-shot human-AI coordination. *arXiv preprint arXiv:2112.11701*.