Loss of Plasticity: A New Perspective on Solving Multi-Agent Exploration for Sparse Reward Tasks

Zehua Zang
University of Chinese Academy of Sciences
Institute of Software Chinese Academy of Sciences
Beijing, China
zehua2020@iscas.ac.cn

Chuxiong Sun*
Institute of Software Chinese Academy of Sciences
Beijing, China
chuxiong2016@iscas.ac.cn

Lixiang Liu
Institute of Software Chinese Academy of Sciences
Beijing, China
lixiang@iscas.ac.cn

Fuchun Sun
Tsinghua University
Institute of Software Chinese Academy of Sciences
Beijing, China
fcsun@tsinghua.edu.cn

Changwen Zheng
Institute of Software Chinese Academy of Sciences
Beijing, China
changwen@iscas.ac.cn

ABSTRACT

Exploration remains a fundamental yet challenging problem in Multi-Agent Reinforcement Learning (MARL). In this paper, we address the issue from a novel perspective: **the loss of plasticity**, a phenomenon characterized by the declining adaptability of neural networks to adapt to new trajectories as training progresses. Through systematic empirical studies, we derive several key insights: (1) Plasticity loss is widespread in MARL; (2) Without timely interventions to restore plasticity, neural networks struggle to learn effective exploration strategies, even when provided with novel and informative data; (3) While restoring plasticity can enhance learning capabilities and exploration efficiency, the process is inherently unstable, with its effectiveness largely depending on which modules are restored and the timing of the intervention. Based on these findings, we propose Plasticity-Aware Multi-Agent Exploration (PAME), which introduces targeted and minimal interventions to enhance plasticity in specific modules of MARL at optimal times. Our results show that PAME consistently outperforms state-of-the-art methods in terms of exploration efficiency.

KEYWORDS

Multi-Agent Reinforcement Learning; Multi-Agent Exploration; Plasticity Loss; Sparse Reward

ACM Reference Format:
Zehua Zang, Chuxiong Sun, Lixiang Liu, Fuchun Sun, and Changwen Zheng. 2025. Loss of Plasticity: A New Perspective on Solving Multi-Agent Exploration for Sparse Reward Tasks. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

*Corresponding author.

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

1 INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) holds significant potential for addressing various real-world multi-agent challenges, such as sensor networks [47], traffic control [37], cooperative robotic systems [11, 44], and Game AI [32]. These complex applications present two primary challenges for cooperative MARL: scalability, due to the exponential growth of the joint action space as the number of agents increases, and partial observability, which necessitates that agents make decentralized decisions based on local action-observation histories. Centralized Training with Decentralized Execution (CTDE) [23] provides an effective solution by leveraging centralized access to global information during training while enabling decentralized decision-making based solely on local observations. CTDE not only ensures scalability for large-scale tasks through decentralized execution but also stabilizes the learning process during centralized training, thereby effectively mitigating the non-stationarity caused by partial observability.

Building upon the CTDE paradigm, value factorization methods [33, 39–43, 45, 46], which employ neural networks to represent the joint Q-value as a function of individual Q-value functions, have achieved considerable success. However, exploration remains a fundamental challenge in MARL, particularly in sparse reward settings, due to the exponential growth of the joint exploration space. Existing exploration strategies in MARL can be broadly categorized into three main approaches, uncertainty-oriented exploration and intrinsic motivated exploration. These methods leverage key insights from exploration techniques in single-agent domains.

Despite recent advances, we observe an intriguing phenomenon: existing multi-agent exploration methods are generally effective at exploring novel states, generating diverse trajectories, and discovering data that can achieve cooperative goals, thereby navigating the state space effectively. However, in complex tasks with sparse rewards, these methods often struggle to translate meaningful exploration data into effective exploration strategies. To address this issue, we revisit the current multi-agent exploration paradigm and

argue that the core challenge behind the low exploration efficiency in policy space is plasticity loss [8].

As the old proverb goes, 'You cannot teach an old dog new tricks.' While traditional wisdom is not always definitive, neuroscientists have long recognized that biological agents gradually lose adaptability with age [22]. This phenomenon, known as plasticity loss, occurs for various reasons, including the natural degradation of neurons and their connections [15, 27]. Recent studies suggest that reinforcement learning (RL) agents utilizing neural networks may similarly lose their ability to learn from new experiences over time [8, 24]. However, understanding the precise mechanisms by which plasticity loss affects exploration efficiency in MARL remains a significant challenge.

In this work, we begin by demonstrating the existence of plasticity loss in MARL. To assess the plasticity of neural networks, we use the number of saturated rectified linear units [24, 38] in a high-performing MARL algorithm, QMIX [33], applied to a variation of the SMAC environment. Our results show a significant decline in the number of saturated rectified linear units over time. We further investigate the degree of plasticity loss across various components of MARL, including the mixer and local Q-value networks in value decomposition methods, as well as the actor and critic networks in multi-agent policy gradient methods. The results indicate that different modules experience varying degrees of plasticity loss. We then analyze how plasticity loss affects exploration efficiency in MARL and find that, when a neural network loses its plasticity, it fails to convert high-quality exploration data into effective exploration strategies, even when such data is available. Finally, we explore how restoring neural network plasticity can enhance exploration efficiency in MARL by conducting motivational experiments to evaluate the impact of restoring plasticity timing on MARL performance.

Building on these insights, we propose a simple yet effective method called Plasticity-Aware Multi-Agent Exploration (PMAE). PMAE not only focuses on exploring novel states and trajectories but also aims to effectively transfer the explored data into a stable exploration policy. The key insight of PMAE is to maintain plasticity throughout the exploration process, ensuring that the knowledge embedded in the exploration data can be successfully integrated into the exploration strategy. Specifically, we leverage Random Network Distillation (RND) on global states to model the intrinsic reward. This intrinsic reward serves a dual purpose: encouraging multi-agent exploration to discover novel states and assessing the novelty of the explored data to determine when to maintain or restore plasticity. To further facilitate this, we introduce a plasticity restoration mechanism for multi-agent exploration that intervenes to enhance the plasticity of the agents' specific neural networks. The conceptual approach is straightforward: at a point when agents have explored novel data, we freeze the current network and create a new one that learns changes to the predictions, while ensuring that these changes initially have no impact. Importantly, the plasticity restoration mechanism does not increase the number of trainable parameters and does not affect the network's predictions when applied. This mechanism ensures that, when agents discover valuable data, the neural network can effectively learn from the trajectory data, thereby improving exploration efficiency in MARL. We evaluate PMAE across a variety of MARL environments, including Google Research Football (GRF) [18] and the StarCraft Multi-Agent Challenge (SMAC) [35]. In both environments, we consider sparse-reward settings, which are particularly challenging since agents must coordinate their behavior over extended timesteps before receiving any non-zero reward. The results demonstrate that PMAE consistently outperforms state-of-the-art baselines, such as ICES [21], EMC [49], CDS [19], SMMAE [48], and COIN [20].

2 RELATED WORKS

Multi-agent exploration can be categorized into two primary approaches based on distinct conceptual foundations. The first approach is uncertainty-oriented exploration, which is rooted in the Optimism in the Face of Uncertainty (OFU) principle. The second approach, intrinsic motivation-oriented exploration, is inspired by the concept of intrinsic motivation in psychology [3], where exploration is guided by intrinsic rewards designed to foster exploratory behavior.

Uncertainty-oriented Exploration Uncertainty-driven exploration faces two key challenges: the agent's reliance on local observations for partial state estimates and the lack of access to other agents' policy, resulting in a non-stationary environment. MSQA [50] models the posterior of the Q-function using a Gaussian process. [28] measures both aleatoric and epistemic uncertainty to guide exploration. [2, 14, 34] extend exploration strategies to zero-sum stochastic games. Their findings indicate that Thomson sampling and Bayes-UCB-based methods are the most effective approaches.

Intrinsic motivation-oriented Exploration Intrinsic motivation-oriented exploration is adding a exploration bonues to the reward signal to encourage agents exploring unseen stats. [4, 12] assign agents extra bonuses based on novelty to encourage exploration. LIIR [10] is proposed which learns the individual intrinsic reward and uses it to update an agent's policy with the objective of maximizing the team reward. [13] define the intrinsic reward function from another perspective called "social influence", which measures the influence of one agent's actions on others' behavior. [6] tackle the coordinated exploration problem from a different view by considering that the environment dynamics caused by joint actions are different from that caused by individually sequential actions.

2.1 Loss of Plasticity

Recent studies have increasingly underscored a significant limitation in neural networks, where their learning capabilities experience catastrophic degradation when trained on non-stationary objectives [29, 38]. Unlike supervised learning, nonstationarity in data streams and optimization objectives is intrinsic to the reinforcement learning (RL) paradigm, making it necessary to address this issue. This challenge has been referred to by several terms, including primacy bias [31], dormant neuron phenomenon [38], implicit underparameterization [16], capacity loss [25], and more broadly, plasticity loss [17, 26]. Agents suffering from plasticity loss struggle to learn from new experiences, resulting in extreme sample inefficiency or, in some cases, entirely ineffective training. The most straightforward approach to mitigate this problem involves reinitializing a portion of the network to restore its plasticity [9, 31, 36].

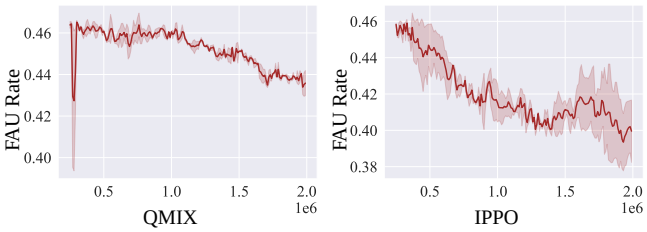

Figure 1: Training curves of FAU rates by training 2M time steps on *3s_5z_vs_3s_6z*.

However, periodic resetting [31] can lead to abrupt performance declines, disrupting exploration and necessitating extensive gradient updates for recovery. To address this limitation, methods such as ReDo [38], which selectively resets dormant neurons, and Plasticity Injection [29], which introduces a new initialized network for learning while freezing the existing network as residual blocks, have been proposed. Another avenue of research emphasizes the use of explicit regularization or modifications to network architecture to counteract plasticity loss. For instance, L2-Init [17] regularizes the network's weights back to their initial parameters, while Concatenated ReLU [1] ensures a non-zero gradient. To the best of our knowledge, we are the first to investigate the problem of plasticity loss and its effects on exploration efficiency in the context of multi-agent reinforcement learning (MARL).

3 DOES PLASTICITY LOSS EXIST IN MARL?

How to evaluate the plasticity in MARL. Although the complete mechanisms behind plasticity loss are still not fully understood, one of the primary contributing factors is a reduction in the number of active neurons within the network. To evaluate plasticity loss in MARL, we employ the Fraction of Active Units (FAU) [24, 38] as a metric to quantify the level of plasticity. The FAU provides insight into the proportion of neurons actively responding to input, offering a direct measure of how effectively the network remains capable of learning. Specifically, the FAU for neurons within MARL networks is defined as follows:

$$\Phi_{\mathcal{M}} = \frac{\sum_{n \in \mathcal{M}} \mathbf{1}(a_n(x) > 0)}{N}, \tag{1}$$

where $a_n(x)$ represents the activation value of neuron n given input x, $\mathcal{M}$ refers to the neural network modules involved in MARL, and N is the total number of neurons within module $\mathcal{M}$. Essentially, $\Phi_{\mathcal{M}}$ captures the ratio of neurons that are actively contributing to the learning process, helping us identify how network plasticity evolves over time.

To demonstrate the existence of plasticity loss in MARL, we selected a representative MARL environment SMAC and chose a challenging task *3s_5z_vs_3s_6z* with high exploration and learning difficulty. We then tracked the changes in FAU during the training processes of two classical MARL algorithms, QMIX [33] and IPPO [7]. As shown in Figure 1, we observe that, as training progresses, the FAU significantly decreases in both Q-learning-based and policy gradient-based MARL algorithms. Despite differences in network architecture, the decline in FAU is evident, indicating that plasticity loss is a widespread issue in the MARL domain.

Table 1: Statistical information by training 60000 time steps on *push box*.Exploration success indicates whether the agent have finished the task while exploration. Successful episodes indicates how many episodes have the agent finished the task while exploration. Win rate indicates the test won rate after training.

Methods	Exploration Success	Successful Episodes	Win Rate
QMIX	N	0	0%
CDS	Y	12	0
COIN	N	0	0
EMC	Y	153	5.2%
ICES	Y	233	3.2%
SMMAE	Y	72	0
RND	Y	201	4.8%

In the following sections, we will discuss whether plasticity loss affects learning and exploration performance in MARL, and if so, how this issue can be addressed.

4 HOW DOES PLASTICITY LOSS AFFECT EXPLORATION EFFICIENCY IN MARL?

Since both the ability of agents to explore new data and the capacity of neural networks to efficiently learn from this data are crucial for the learning efficiency of MARL—both being indispensable—understanding the influence of plasticity on exploration efficiency in MARL first requires an investigation into the respective roles of these two aspects in MARL tasks where exploration fails. Hence, we consider a Push-Box task, where two agents need to jointly push a heavy box to a specific location before observing a reward. In this task, the exploration of agents position is easy but exploration of box's position require agents' cooperative efforts. Furthermore, until pushing the box within the environment to find the specific location, agents can not get reward signal, therefore many existing state-of-the-art methods fail to explore in this task. As shown in Table 1, we consider trajectories that complete the task as successful explorations, and use these successful trajectories to assess whether agents have discovered effective data. The results indicate that not all learning failures are due to a lack of exploration. On the contrary, most multi-agent exploration methods are able to find effective data even in complex tasks with sparse rewards. However, despite discovering such data, the converged win rate remains at zero, demonstrating an inability to learn a stable exploration strategy from effective data.

To further illustrate this phenomenon, we collected an offline dataset containing of numerous successful exploration trajectories. Using this dataset, we construct two policy networks exhibiting different levels of plasticity: one is a randomly initialized network, representing a high degree of plasticity, while the other is a network that had undergone extensive training, leading to a significant reduction in the number of saturated rectified linear units, thereby representing low plasticity. As shown in Figure 3, the highly plastic network is able to quickly fit an effective exploration strategy from the high-quality offline dataset, rapidly discovering a solution that could reliably solve the task. In contrast, the low-plasticity network

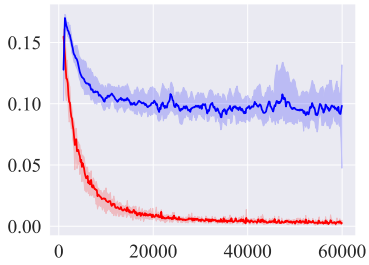

Figure 2: Training curves of losses by training 60000 time steps on *push box*. The blue curve is trained with low-plasticity network. The red curve is trained with highly plastic network.

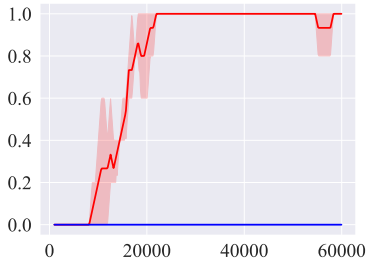

Figure 3: Training curves of rewards by training 60000 time steps on *push box*. The blue curve is trained with low-plasticity network. The red curve is trained with highly plastic network.

struggled to learn an effective strategy, with its win rate remaining close to zero throughout the training process. At the same time, the training losses of the highly plastic network decreases faster than that of low-plasticity network, as shown in Figure 2. These results further highlight the impact of plasticity loss on MARL learning efficiency. When studying the multi-agent exploration problem, it is crucial not only to consider whether novel data can be explored but also whether the current policy network can effectively learn from that novel data.

5 HOW TO RECOVER PLASTICITY AND IMPROVE EXPLORATION EFFICIENCY IN MARL?

In previous sections, we have demonstrated that plasticity loss is prevalent in MARL and significantly affects exploration efficiency. In this section, we aim to address the following questions:

How can plasticity be recovered? A straightforward approach to recover plasticity is to reinitialize parts of the network, thereby rejuvenating its ability to learn. However, this strategy can disrupt the exploration process and requires numerous gradient updates to regain the lost performance, making it costly and inefficient.

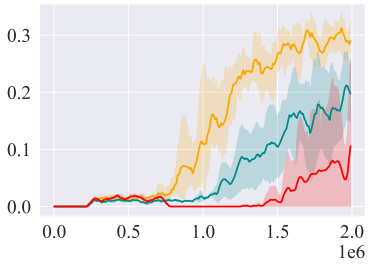

Figure 4: Training curves of test battle won mean by 2M steps on *3s_5z_vs_3s_6z*. Yellow curves indicates that only recover the plasticity of mixer network. Dark cyan curves indicates that only recover the plasticity of local agent network. Red curves indicates that recover the plasticity of all the networks.

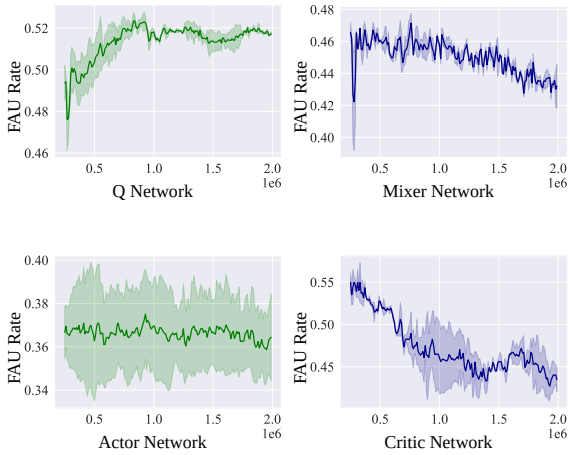

Figure 5: Training curves of FAU rates by 2M steps on *3s_5z_vs_3s_6z*.

Which components should be targeted for recovery? Our motivation experiment, presented in Figure 4, confirms this drawback: the performance of a complete plasticity reset across the entire MARL network is significantly lower compared to selectively restoring plasticity in specific, critical modules. Therefore, it is essential to determine which modules within the MARL network are most severely impacted by plasticity loss and selectively restore their plasticity. By doing so, we can maintain training stability while effectively recovering the network's plasticity, minimizing disruption and optimizing the learning process. To further investigate this, we evaluated the FAU during the learning process of both QMIX and IPPO. Specifically, we assessed the local Q-value network and the mixer network in QMIX, as well as the actor and critic networks in IPPO. As shown in Figure 5, our findings indicate the following: (1) In QMIX, the mixer network is significantly more affected by plasticity loss compared to the local Q-value network; (2) In IPPO, the critic network experiences more severe plasticity loss than the actor network.

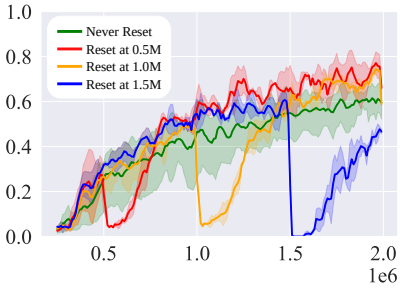

Figure 6: Training curves of by 2M steps on *3s_5z_vs_3s_6z*. Different curves indicate reset at a different training stages.

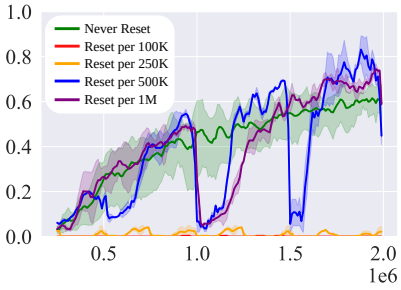

Figure 7: Training curves of by 2M steps on *3s_5z_vs_3s_6z*. Different curves indicate reset with a specific frequency.

When is the optimal time to perform recovery? Upon confirming that plasticity loss in specific MARL modules is a key factor hampering training, we further investigate the impact of restoring plasticity at different stages of training and with different frequency. As shown in Figure 6 and Figure 7, we find that restoring plasticity at different stages results in substantial variations in the performance of the learned policies. Therefore, determining an appropriate timing for stable plasticity restoration is an important challenge that needs to be addressed. In this work, we study the relationship between plasticity and exploration, concluding that the two are mutually reinforcing and indispensable for effective learning. Based on this insight, we hypothesize that when agents discover high-quality novel data through exploration, the neural network must retain sufficient plasticity to effectively learn from this data. Hence, we propose a plasticity-aware multi-agent exploration method. The detailed methods will be introduced in the following section.

6 METHODS: PLASTICITY-AWARE MULTI-AGENT EXPLORATION

In this section, we mainly focus on answering the core question discussed above: How to learn effective policy based on successful exploration from the perspective of plasticity. Thus, we divide this section into four subsections: the first subsection is the preliminaries for introducing the background of our method, the second subsection is how to conduct successful exploration, the third subsection is how to restore the plasticity of agents, and the fourth subsection is when to recover the plasticity of agents.

6.1 Preliminaries

We focuses on fully cooperative multi-agent reinforcement learning tasks, characterized by partial observability. These tasks build upon the framework of Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs), formalized through the tuple $G = (N, S, O, A.R, \gamma, P)$. In this formulation, $N = (agent_1, \ldots, agent_n)$ represents the set of agents. S refers to the global states, which provide a complete representation of the environment. O denotes the local observations accessible to the agents, while A defines the available actions. The reward function R is based on the global states and joint actions, and γ represents the discount factor. The transition function P governs the environmental dynamics.

6.2 Explore Novel Experiences by Prediction Error

The core premise of our method centers on the acquisition of novel experiences. If novel experiences cannot be obtained, the failure to solve tasks can be attributed to inadequate exploration, rather than the inability to learn from novel data as discussed in this paper. A widely adopted strategy to enhance exploration is to incentivize agents to engage in more diverse interactions through the use of intrinsic rewards. Drawing inspiration from the prediction error in RND [5], a well-established exploration method in single-agent reinforcement learning, we utilize prediction error as novelty-based intrinsic rewards to promote exploration. The fundamental idea is to assign higher intrinsic rewards to states the agent has rarely encountered, thereby encouraging exploration of less familiar regions of the environment. At the beginning of training, a target network, $\phi^\star$, with fixed parameters is randomly initialized to serve as an anchor, while a predictor network, ϕ, is trained to approximate the output of the target network.

Given a state s_t at time step t, both the target network and the predictor network take s_t as input and produce corresponding embeddings. The output of the target network, $f_{\phi^\star}(s_t)$, remains fixed as the target network is not trained, while the output of the predictor network, $f_\phi(s_t)$, is optimized to approximate that of the target network. The intrinsic reward is determined by the prediction error between the outputs of the two networks. Formally, the intrinsic reward at time step t, denoted as r_t^{int}, is computed as the squared error:

$$r_t^{\text{int}} = \|f_{\phi^\star}(s_t) - f_\phi(s_t)\|^2 \tag{2}$$

A high prediction error signifies that the current state, s_t, is novel or less familiar, as the predictor network has not yet learned to replicate the target network's output for that state. The agent receives this intrinsic reward in addition to any extrinsic reward provided by the environment. Since novel states result in larger prediction errors, the agent is encouraged to explore less-visited regions of the state space where the predictor network exhibits higher errors. Over time, as the predictor network is trained on a broader range of states, its performance improves, leading to a reduction in intrinsic rewards for states that were previously novel. This dynamic incentivizes the agent to continually explore new areas in order to maintain high intrinsic rewards. Ultimately, the total reward used for training is the sum of the extrinsic reward,

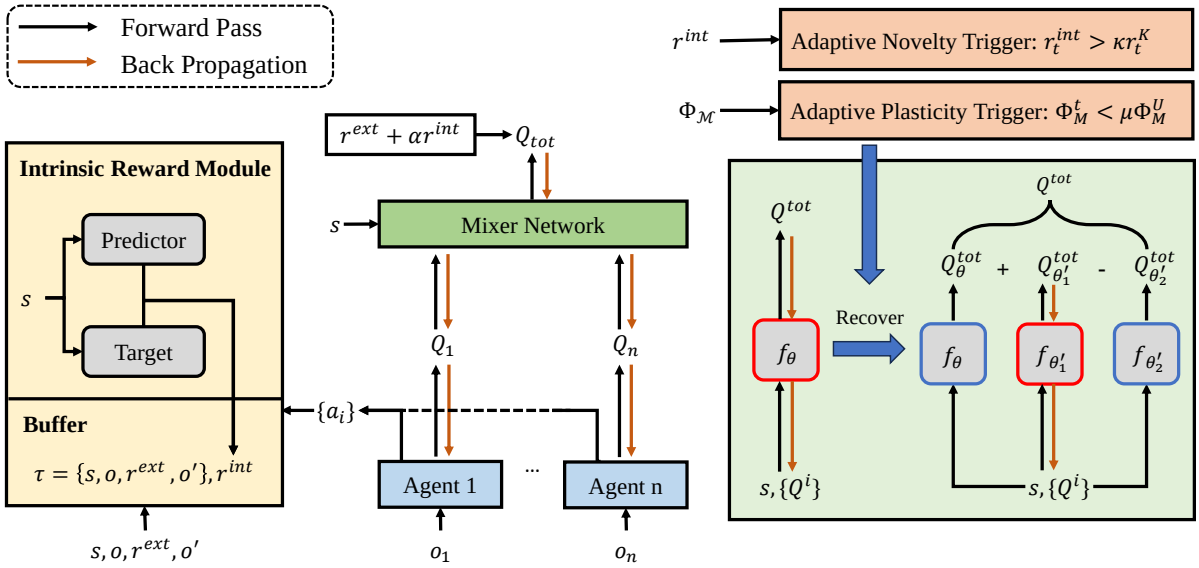

Figure 8: Overall framework of PAME., At each timestep t, agent i receive an observation o_i, and generate an action a_i. The intrinsic reward r^{int} is calculated by the intrinsic reward module and used to optimize mixed Q value with extra reward r^{ext}. At the same time, when both trigger are triggered, the plasticity of specific network module is recovered by $f_\theta = f_\theta - f_{\theta'_1} + f_{\theta'_2}$.

r_t^{ext}, from the environment and the intrinsic reward, r_t^{int}:

$$r_t^{total} = r_t^{ext} + \alpha r_t^{int} \tag{3}$$

where α is a hyper-parameter used to balance external reward and intrinsic reward.

6.3 Recover Plasticity of Modules

In addition to effective exploration, an agent's ability to learn from novel experiences is a critical factor for solving tasks. When an agent is unable to derive effective policies from novel experiences due to a loss of plasticity, mechanisms to restore plasticity become more essential than simply continuing to explore new experiences. By leveraging Plasticity Injection [30], a method that enhances plasticity without affecting predictions or increasing the trainable parameter count, we introduce new trainable networks to restore the plasticity of the agent. At a certain point during training, when the plasticity of a module θ has already degraded, the parameters of θ are frozen, and two randomly initialized networks, θ'_1 and θ'_2, are introduced. The network θ'_1 contains trainable parameters used to learn a residual to the outputs of the original network, while θ'_2 remains frozen throughout. Given an input x, the predictions of the recovered agent are computed as follows:

$$f(x) = f_\theta(x) + f_{\theta'_1}(x) - f_{\theta'_2}(x) \tag{4}$$

Since $\theta'_1 = \theta'_2$ at initialization, the agent's previously learned knowledge remains unaffected during the plasticity recovering process. As learning progresses, θ'_1 diverges from θ'_2, and the difference $f_\theta(x) - f_{\theta'_2}(x)$ functions as a bias term in the predictions. In practice, it is not necessary to recover all modules. As discussed in Section 5, applying recovering to all the modules of agents forces the agent to relearn its entire representation from scratch which is inefficient to training, so we only target modules where plasticity has significantly diminished, such as the mixer network in QMIX [33]-based methods or the critic network in Actor-Critic methods.

6.4 Adaptive Triggers for Recovering

Recovering plasticity is an effective strategy for enhancing an agent's ability to learn policies from novel experiences. However, when to recover the plasticity is equally critical. For instance, recovering at initialization stage or at the final stage of training would yield the same outcome as not recovering at all. To address this, we propose two adaptive triggers for deciding a suitable time to recover plasticity. Before outlining the experimental design, we first highlight the key motivating criteria:

(1) **Novel experiences:** The experiences in the replay buffer must be sufficiently novel. This ensures that recovering occurs in the context of effective exploration.
(2) **Plasticity loss:** Recovering is applied only when the agent's plasticity has significantly diminished.

We now introduce the proposed adaptive triggers to recover the agent's plasticity. On one hand, it is essential to ensure that when the recovering trigger is activated, the experiences in the buffer are novel enough. Therefore, the novelty of experiences in the replay buffer $\mathcal{D}$ must be measured. Fortunately, during the exploration phase, we have already established indicators of experience novelty. Thus, we continue to use these previously defined evaluation metrics and compute the sum of predicted losses across all experiences in the buffer to assess the overall novelty r_t^{int} at time step t:

$$r_t^{int} = \sum_{s \sim \mathcal{D}} \| f_{\phi^\star}(s) - f_\phi(s) \|^2 \tag{5}$$

Specifically, we record the novelty of an anchor buffer at time step K, where K is a small integer greater than zero. The first trigger, denoted as adaptive novelty trigger, is set as $\eta_t > \kappa\eta_K$ with $\kappa > 1$, meaning that recovering the module's plasticity when the novelty at time step t exceeds a scaled version of the novelty at step K. The rationale for selecting K as a small integer greater than zero is that the buffer requires time to populate. And as the predictor network is trained, the overall novelty of the buffer, which includes repeated experiences, decreases. A lower value of η_K makes the trigger easier to activate. The parameter κ introduces a trade-off: if κ is too small, recovering may occur prematurely, akin to a random initialization reset. Conversely, if κ is too large, the agent may find it difficult to explore sufficiently novel experiences to trigger recovering.

In addition, recovering plasticity must be applied to modules that exhibit plasticity loss. To measure plasticity, we propose using FAU as a metric. Following Equation 1, we record the FAU, denoted $\Phi^U_{\mathcal{M}}$, for an anchor module at time step U, where U is a small integer greater than zero. The second trigger, which is called adaptive plasticity trigger, is set as $\Phi^t_{\mathcal{M}} < \mu\Phi^U_{\mathcal{M}}$, with $0 < \mu < 1$. The reason for selecting U as a small integer greater than zero is that FAU is highly unstable at the beginning of training and tends to decrease as the module is trained. The parameter μ introduces a trade-off: if μ is too small, the trigger is never activated, while if μ is too large, recovering occurs prematurely.

7 EXPERIMENTS

7.1 Experimental Settings

In this work, we evaluate PAME and baselines on widely used benchmarks of GRF and SMAC in **sparse** reward settings. We consider three tasks in GRF: *3_vs_1_with_keeper*, *corner* and *counterattack_hard*, and five tasks in SMAC: *2c_vs_64zg*, *5m_vs_6m*, *8m_vs_9m*, *MMM* and *MMM2*. We implement our proposed PAME on top of QMIX [33]. We compare PAME with CDS [19], COIN [20], EMC [49], ICES [21] and SMMAE [48]. Wherever possible, we utilize the official implementations of these baselines from their respective papers; in cases where the implementation is not available, we closely follow the descriptions provided in the papers and implement them on top of QMIX.

7.2 Benchmark Results on GRF and SMAC

Figure 9 compares the performance of PAME and several baseline methods across GRF and SMAC tasks under sparse reward settings. PAME consistently outperforms the baselines, especially those based on QMIX, demonstrating its ability to learn effectively in challenging, sparse-reward environments. PAME shows a sharp performance increase after a certain number of time steps, indicating efficient exploration and policy learning, while baseline methods either struggle to improve or exhibit slower progress. This suggests that PAME enhances the agents' ability to learn from novel experiences when the recovery triggers, as discussed in Section 6.4, are met.

In the GRF tasks, which are known for their environmental stochasticity and high demand for agent collaboration, PAME demonstrates its robustness. For instance, in the *3_vs_1_with_keeper* environment, PAME performs on par with the baselines up until approximately 1.2 million steps. However, after this point, a significant performance improvement occurs, signaling the potential influence of plasticity recovering around this time. Similar trends are observed in the *corner* and *counterattack_hard* environments.

Turning to the SMAC tasks, PAME similarly exhibits superior performance compared to the baselines. In tasks such as *5m_vs_6m*, *8m_vs_9m*, and *MMM2*, PAME achieves marked performance improvements at specific time steps, while the baselines show only marginal gains or even fail to solve the tasks. This supports our earlier conclusion that successful exploration not only requires encountering novel experiences but also the capacity to learn from them. In the remaining tasks, *2c_vs_64zg* and *MMM*, both PAME and the baselines successfully solve the tasks, but PAME demonstrates faster convergence. This suggests that while the baseline methods suffer from some degree of plasticity loss, they do not entirely lose the ability to learn. In contrast, PAME accelerates learning by restoring the agent's plasticity, enabling quicker adaptation to the problem-solving requirements.

7.3 Ablation Studies

In this subsection, we further investigate the effectiveness of our two proposed triggers and the exploration mechanism based on prediction error as intrinsic rewards. We evaluated the following settings, recover nothing which is only exploration with intrinsic rewards (Recover Nothing), recover mixer network (Recover Mixer), recover local agents' Q network (Recover Agent) and recover all the networks (Recover All), as shown in Figure 10. It should be noticed that the time of plasticity recovery in all the evaluations is at 0.75 million time steps. These evaluations are based on 2M steps training on 5m_vs_6m.

Ablation Study on Adaptive Novelty Trigger. The adaptive novelty trigger plays a critical role in determining whether the condition for plasticity recovery based on experience novelty is met. To assess the impact of recovery timing on final performance, we compare the performance curves of PAME and Recover Mixer. The key difference between these two approaches is that PAME employs the proposed adaptive novelty trigger, while Recover Mixer utilizes a fixed recovery time of 0.75 million time steps, which is later than the adaptive trigger.

The results demonstrate that PAME achieves faster performance improvements compared to Recover Mixer, although both methods exhibit similar overall performance trends. The plasticity of the mixer network declines rapidly during training, and restoring plasticity leads to substantial improvements in the network's ability to learn from novel experiences. However, the learning capacity of the mixer network after plasticity recovery appears consistent, regardless of when the recovery occurs. Therefore, within an appropriate range, earlier plasticity recovery enables the agent to achieve optimal performance more rapidly. This finding underscores the importance of timely plasticity recovering in accelerating the agent's learning process.

Ablation Study on Adaptive Plasticity Trigger. The adaptive plasticity trigger determines if recovery is needed based on plasticity loss, preventing unnecessary recovery for unaffected modules. To validate this, we compare the performance of four approaches—Recover Nothing, Recover Mixer, Recover Agent, and

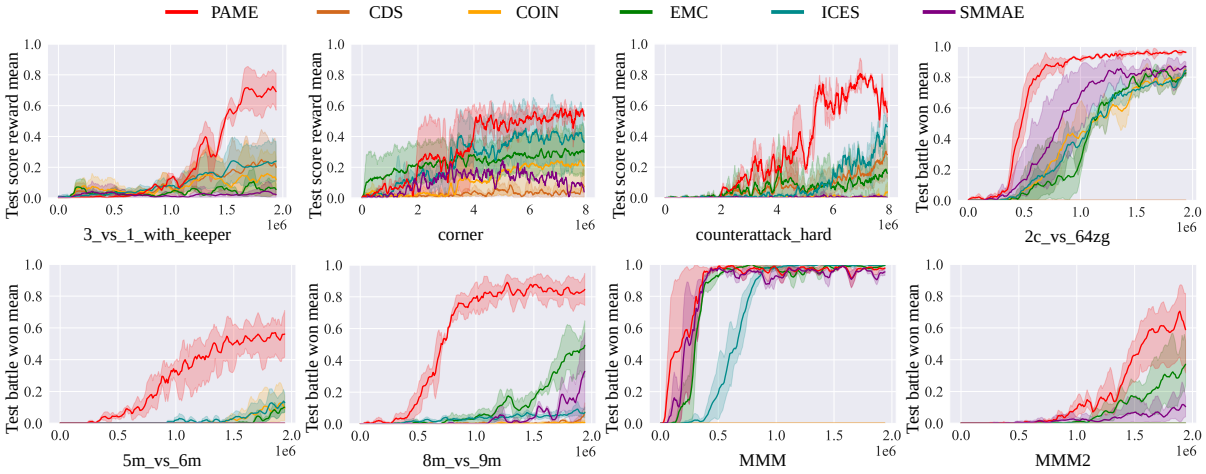

Figure 9: Performance comparison with baselines on GRF and SMAC benchmarks in sparse reward settings.

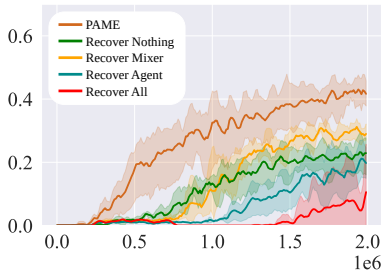

Figure 10: Training curves of by 2M steps on *5m_vs_6m*.

Recover All—which all start recovery at 0.75 million time steps but with different recovery targets.

The results show that the optimal strategy is to recover only the mixer network's plasticity, with Recover Mixer outperforming the others. Recover Nothing, which applies no recovery, is the second-best, followed by Recover Agent, which recovers the Q network of local agents. Recover All, where all modules undergo recovery, performs the worst. As discussed in Section 5, selectively recovering plasticity from modules with significant loss is beneficial, while unnecessary recovery harms performance. After 0.75 million steps, performance stagnates for about 0.5 million steps before improving, suggesting that recovering plasticity in unaffected modules forces the network to relearn acquired knowledge. This issue is worse in Recover All, where the agent must relearn its entire representation, severely hindering progress. These results support our claim that only modules with substantial plasticity loss should undergo recovery, with the adaptive plasticity trigger ensuring efficient, selective recovery.

Ablation Studies on Hyperparameters To determine the optimal values for the introduced hyperparameters, we conducted three ablation experiments on *5m_vs_6m*, focusing on the intrinsic reward (α), adaptive novelty trigger (κ), and adaptive plasticity trigger (μ). The results, presented in Table 2, show that we set $\alpha = 0.1$, $\kappa = 1.2$, and $\mu = 0.9$ for our experiments.

Table 2: Test battle won rate on *5m_vs_6m*.

α	Won Rate	κ	Won Rate	μ	Won Rate
0	0.19±0.042	1.0	±0.27±0.091	1.0	±0.30±0.107
0.05	0.27±0.078	1.2	**±0.56±0.102**	0.9	**±0.56±0.102**
0.1	**0.41±0.098**	1.4	±0.30±0.020	0.8	±0.28±0.060
0.2	0.32±0.050	1.6	±0.28±0.025	-	-

Effective Exploration Finally, we were curious to see if our work on plasticity recovery had been successful rather than successful in other ways, and when we compared the baselines in Figure 9 and Recover Nothing in Figure 10, we found that our exploration of predicting loss as an intrinsic reward was successful in exploring novel experiences.

8 CONCLUSIONS

In conclusion, our findings highlight the significant impact of plasticity loss on exploration efficiency in MARL. By identifying this degradation as a critical barrier to learning, we emphasize the need for timely plasticity restoration. The proposed Plasticity-Aware Multi-Agent Exploration (PAME) method offers a lightweight yet effective intervention to address this challenge, achieving consistent improvements in exploration without increasing model complexity. These results suggest that maintaining plasticity throughout training is essential for advancing exploration strategies in MARL.

ACKNOWLEDGMENT

This work is supported by the Fundamental Research Program, China, Grant No. JCKY2022130C020, National Natural Science Foundation of China, Grant No. 62406313, Postdoctoral Fellowship Program, Grant No. GZC20232812, China Postdoctoral Science Foundation, Grant No. 2024M753356, 2023 Special Research Assistant Grant Project of the Chinese Academy of Sciences.

REFERENCES

[1] Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C. Machado. 2023. Loss of Plasticity in Continual Deep Reinforcement Learning. In *Conference on Lifelong Learning Agents, 22-25 August 2023, McGill University, Montréal, Québec, Canada (Proceedings of Machine Learning Research, Vol. 232)*, Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina Precup (Eds.). PMLR, 620–636. https://proceedings.mlr.press/v232/abbas23a.html

[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* 47, 2-3 (2002), 235–256. https://doi.org/10.1023/A:1013689704352

[3] Andrew G. Barto. 2013. Intrinsic Motivation and Reinforcement Learning. In *Intrinsically Motivated Learning in Natural and Artificial Systems*, Gianluca Baldassarre and Marco Mirolli (Eds.). Springer, 17–47. https://doi.org/10.1007/978-3-642-32375-1_2

[4] Wendelin Böhmer, Tabish Rashid, and Shimon Whiteson. 2019. Exploration with Unreliable Intrinsic Reward in Multi-Agent Reinforcement Learning. *CoRR* abs/1906.02138 (2019). arXiv:1906.02138 http://arxiv.org/abs/1906.02138

[5] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* (2018).

[6] Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. 2020. Intrinsic Motivation for Encouraging Synergistic Behavior. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SJleNCNtDH

[7] Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533* (2020).

[8] Shibhansh Dohare, Richard S Sutton, and A Rupam Mahmood. 2021. Continual backprop: Stochastic gradient descent with persistent randomness. *arXiv preprint arXiv:2108.06325* (2021).

[9] Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G. Bellemare, and Aaron C. Courville. 2023. Sample-Efficient Reinforcement Learning by Breaking the Replay Ratio Barrier. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=OpC-9aBBVJe

[10] Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. 2019. LIIR: Learning Individual Intrinsic Reward in Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 4405–4416. https://proceedings.neurips.cc/paper/2019/hash/07a9d3fed4c5ea6b17e80258dee231fa-Abstract.html

[11] Maximilian Hüttenrauch, Adrian Šošić, and Gerhard Neumann. 2017. Guided deep reinforcement learning for swarm systems. *arXiv preprint arXiv:1709.06011* (2017).

[12] Shariq Iqbal and Fei Sha. 2019. Coordinated Exploration via Intrinsic Rewards for Multi-Agent Reinforcement Learning. *CoRR* abs/1905.12127 (2019). arXiv:1905.12127 http://arxiv.org/abs/1905.12127

[13] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3040–3049. http://proceedings.mlr.press/v97/jaques19a.html

[14] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2012. On Bayesian Upper Confidence Bounds for Bandit Problems. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012 (JMLR Proceedings, Vol. 22)*, Neil D. Lawrence and Mark A. Girolami (Eds.). JMLR.org, 592–600. http://proceedings.mlr.press/v22/kaufmann12.html

[15] Bryan Kolb and Robbin Gibb. 2011. Brain plasticity and behaviour in the developing brain. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 20, 4 (2011), 265.

[16] Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. 2021. Implicit Under-Parameterization Inhibits Data-Efficient Deep Reinforcement Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=O9bnihsFfXU

[17] Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. 2023. Maintaining Plasticity via Regenerative Regularization. *CoRR* abs/2308.11958 (2023). https://doi.org/10.48550/ARXIV.2308.11958 arXiv:2308.11958

[18] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 4501–4510.

[19] Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 3991–4002.

[20] Jiahui Li, Kun Kuang, Baoxiang Wang, Xingchen Li, Fei Wu, Jun Xiao, and Long Chen. 2024. Two heads are better than one: a simple exploration framework for efficient multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

[21] Xinran Li, Zifan Liu, Shibo Chen, and Jun Zhang. 2024. Individual Contributions as Intrinsic Exploration Scaffolds for Multi-agent Reinforcement Learning. *CoRR* abs/2405.18110 (2024). https://doi.org/10.48550/ARXIV.2405.18110 arXiv:2405.18110

[22] Long-Ji Lin. 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning* 8 (1992), 293–321.

[23] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*. 6379–6390.

[24] Clare Lyle, Mark Rowland, and Will Dabney. 2022. Understanding and preventing capacity loss in reinforcement learning. *arXiv preprint arXiv:2204.09560* (2022).

[25] Clare Lyle, Mark Rowland, and Will Dabney. 2022. Understanding and Preventing Capacity Loss in Reinforcement Learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=ZkC8wKoLbQ7

[26] Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Ávila Pires, Razvan Pascanu, and Will Dabney. 2023. Understanding Plasticity in Neural Networks. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 23190–23211. https://proceedings.mlr.press/v202/lyle23b.html

[27] Henry W Mahncke, Amy Bronstone, and Michael M Merzenich. 2006. Brain plasticity and functional losses in the aged: scientific bases for a novel intervention. *Progress in brain research* 157 (2006), 81–109.

[28] Carlos Martin and Tuomas Sandholm. 2020. Efficient exploration of zero-sum stochastic games. *CoRR* abs/2002.10524 (2020). arXiv:2002.10524 https://arxiv.org/abs/2002.10524

[29] Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and André Barreto. 2023. Deep Reinforcement Learning with Plasticity Injection. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/75101364dc3aa7772d27528ea504472b-Abstract-Conference.html

[30] Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and André Barreto. 2023. Deep Reinforcement Learning with Plasticity Injection. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/75101364dc3aa7772d27528ea504472b-Abstract-Conference.html

[31] Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron C. Courville. 2022. The Primacy Bias in Deep Reinforcement Learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 16828–16847. https://proceedings.mlr.press/v162/nikishin22a.html

[32] Peng Peng, Ying Wen, Yaodong Yang, Quan Yuan, Zhenkun Tang, Haitao Long, and Jun Wang. 2017. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069* (2017).

[33] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485* (2018).

[34] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2018. A Tutorial on Thompson Sampling. *Found. Trends Mach. Learn.* 11, 1 (2018), 1–96. https://doi.org/10.1561/2200000070

[35] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043* (2019).

[36] Max Schwarzer, Johan Samir Obando-Ceron, Aaron C. Courville, Marc G. Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. 2023. Bigger, Better, Faster: Human-level Atari with human-level efficiency. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill,

Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 30365–30380. https://proceedings.mlr.press/v202/schwarzer23a.html

[37] Arambam James Singh, Akshat Kumar, and Hoong Chuin Lau. 2020. Hierarchical Multiagent Reinforcement Learning for Maritime Traffic Management. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, 1278–1286. https://doi.org/10.5555/3398761.3398909

[38] Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. 2023. The Dormant Neuron Phenomenon in Deep Reinforcement Learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 32145–32168. https://proceedings.mlr.press/v202/sokar23a.html

[39] Chuxiong Sun, Peng He, Qirui Ji, Zehua Zang, Jiangmeng Li, Rui Wang, and Wei Wang. 2024. M2I2: Learning Efficient Multi-Agent Communication via Masked State Modeling and Intention Inference. *arXiv preprint arXiv:2501.00312* (2024).

[40] Chuxiong Sun, Peng He, Rui Wang, and Changwen Zheng. 2025. Revisiting Communication Efficiency in Multi-Agent Reinforcement Learning from the Dimensional Analysis Perspective. *arXiv preprint arXiv:2501.02888* (2025).

[41] Chuxiong Sun, Bo Wu, Rui Wang, Xiaohui Hu, Xiaoya Yang, and Cong Cong. 2021. Intrinsic Motivated Multi-Agent Communication. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1668–1670.

[42] Chuxiong Sun, Zehua Zang, Jiabao Li, Jiangmeng Li, Xiao Xu, Rui Wang, and Changwen Zheng. 2024. T2MAC: Targeted and Trusted Multi-Agent Communication through Selective Engagement and Evidence-Driven Integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 15154–15163.

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

[44] Gokul Swamy, Siddharth Reddy, Sergey Levine, and Anca D Dragan. 2020. Scaled autonomy: Enabling human operators to control robot fleets. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5942–5948.

[45] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *ICLR 2021: The Ninth International Conference on Learning Representations*.

[46] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[47] Chongjie Zhang and Victor Lesser. 2011. Coordinated multi-agent reinforcement learning in networked distributed POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25. 764–770.

[48] Shaowei Zhang, Jiahan Cao, Lei Yuan, Yang Yu, and De-Chuan Zhan. 2023. Self-Motivated Multi-Agent Exploration. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh (Eds.). ACM, 476–484. https://doi.org/10.5555/3545946.3598673

[49] Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems* 34 (2021), 3757–3769.

[50] Zheqing Zhu, Erdem Biyik, and Dorsa Sadigh. 2020. Multi-Agent Safe Planning with Gaussian Processes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021*. IEEE, 6260–6267. https://doi.org/10.1109/IROS45743.2020.9341169