When Intrinsic Motivation Fails: Exploration Challenges in Decentralized MARL

Anonymous authors

Abstract

Learning coordinated behaviour in decentralised multi-agent reinforcement learning with sparse rewards presents significant exploration challenges. Whilst novelty bonuses encourage exploration, we identify a critical failure mode they create in sequential coordination tasks: *coordination de-synchronisation*, where agents repeatedly traversing earlier coordination points gradually exhaust their intrinsic motivation to revisit these critical locations. We hypothesize that the effectiveness of exploration strategies depends on two key task factors: coordination complexity and geometric revisit pressure. Our preliminary experiments confirm this: lifelong novelty bonuses deteriorate with increasing task complexity, while augmenting with episodic bonuses substantially improves performance. These findings motivate further theoretical investigation into coordination-aware intrinsic motivation for decentralized agents.

Keywords

Decentralized multi-agent, Reinforcement learning, Exploration, Novelty,

1. Introduction

Cooperative multi-agent reinforcement learning (MARL) has become increasingly important for addressing complex real-world challenges that require coordinated behaviors among multiple agents. Successful applications included wireless sensor networks [1], unmanned aerial vehicles [2], and traffic light control [3]. Unlike single-agent settings, MARL introduces unique challenges—especially in decentralized scenarios where agents must learn to coordinate based solely on local observations, without access to global information. Despite the progress made by recent fully decentralized algorithms [4, 5, 6], a major limitation remains: these methods often struggle in environments with sparse rewards. Exploration becomes especially critical in sparse-reward cooperative tasks, where agents must synchronise their actions at specific locations and times to make progress, yet only receive exceedingly rare external reward signals.

A common approach to exploration in single-agent setting is through intrinsic motivation via novelty-based bonuses—rewards based on how rarely states have been visited [7, 8, 9]. While these methods have been extensively evaluated in single-agent settings, their extension to decentralised multi-agent settings brings unique challenges. Under partial observability, each agent only sees its own local state and thus computes a local novelty estimate - might at odds with the true novelty of the joint state. This gap between local and global novelty measure can misalign exploration incentives across the team. This raises fundamental questions: How should intrinsic motivation be designed for decentralized agents? Can novelty mislead agents away from successful coordination?

In this work, we aim to initiate a discussion around these questions by highlight a critical failure mode - what we term *coordination de-synchronisation*. The core problem arises when cooperative tasks require agents to coordinate at multiple points sequentially. As agents search for later coordination opportunities, they must repeatedly traverse earlier coordination points, gradually exhausting their intrinsic motivation to revisit these critical locations. This creates a coordination ”death spiral” where the very act of exploring for advanced objectives undermines the team’s ability to maintain coordination at foundational checkpoints—precisely the opposite of what we desire from a learning system. Our key insight is that this failure is determined by two independent factors: the task’s *logical complexity* (the number of sequential coordination points) and the environment’s *physical geometry*, which can force agents to repeatedly traverse earlier checkpoints. We demonstrate that a hybrid approach, combining

9th International Workshop on Interactive Adaptive Learning (IAL 2025)

episodic and lifelong bonuses, provides a complete solution. In this work, we present initial empirical results validating our hypothesis, while the accompanying theoretical analysis is ongoing. Our goal is to share early insights and open a broader discussion on coordination-aware exploration design.

2. Related work

Intrinsic motivation in multi-agent RL. Several works have investigated intrinsic motivation in MARL within the CTDE framework. Iqbal and Sha [10] proposed coordinated exploration bonuses that consider novelty from other agents' perspectives, whilst Wang et al. [11] assigned bonuses based on mutual information between agents' transition dynamics. Zheng et al. [12] used prediction errors of individual Q-functions as intrinsic rewards, and Liu et al. [13] addressed lazy agents through causal effect maximisation on external states. These approaches differ fundamentally from our work in two key aspects. First, they operate within CTDE settings where agents have access to global information during training, whilst we focus on the more challenging decentralised case where agents rely solely on local observations. Second, none provide theoretical analysis of when different exploration strategies succeed or fail based on task structure. Jiang et al. [14] introduced MACE, the first intrinsic motivation method specifically for decentralised settings, where agents share only local novelty values. However, MACE focuses on approximating global novelty through local information sharing rather than addressing the coordination challenges we identify. Our work would complement MACE by providing theoretical foundations for when episodic components become crucial in decentralised coordination tasks.

Episodic bonuses in MARL. Toquebiau et al. [15] is the first to combine lifelong and episodic bonuses in MARL, using NovelD [16] and E3B [17] within CTDE. However, they compute novelty using joint observations, giving all agents identical rewards, whilst our decentralised approach requires agents to compute individual novelty from local observations. Critically, their work lacks theoretical analysis of when such combinations become necessary, which is the central contribution of our theoretical framework.

Communication in decentralised MARL. Research on decentralised communication has focused on establishing common grounding between agents. Lin et al. [18] proposed AEComm, where agents autoencode observations for communication using shared environment structure, whilst Lo et al. [19] introduced contrastive learning to align messages from agents perceiving the same global state. Our theoretical analysis focuses on the fundamental case where agents operate independently without communication, revealing how coordination complexity determines exploration strategy effectiveness. While our experiments extend this to include novelty sharing between agents, the core theoretical insight—that episodic bonuses provide coordination robustness while lifelong bonuses suffer decay in sequential tasks—applies regardless of whether agents share novelty information. This suggests the temporal scope of novelty assessment is more fundamental than the communication mechanism itself.

3. Theoretical Framing and Tasks Design

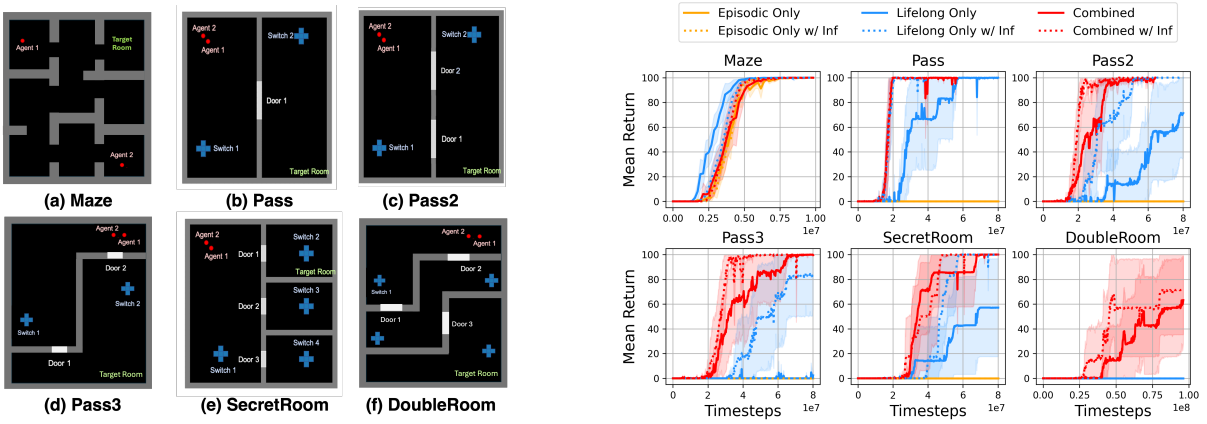
3.1. Fundamental Concepts

Coordination Checkpoints: We study cooperative tasks that require agents to visit an ordered list of *coordination checkpoints*, $S^* = \{s^*_{(1)}, \dots, s^*_{(L)}\}$. Each checkpoint is a joint state $s^*_{(\ell)} = (s^*_{1,\ell}, \dots, s^*_{N,\ell})$ that the team must reach simultaneously. For instance, a simple task might require two agents to stand on two different pressure plates at the same time to open a door. This joint requirement—Agent 1 at position A *and* Agent 2 at position B—constitutes the first checkpoint, $s^*_{(1)}$. This sequence must be fully completed to receive any non-zero extrinsic reward; the reward function R is zero at all times, except for a positive reward granted upon reaching the final checkpoint $s^*_{(L)}$ after all previous checkpoints have been visited in order within the same episode.

Geometric pressure: We define geometric pressure as a property of the environment that reflects how often agents are forced to revisit earlier parts of the map during exploration. High geometric pressure increases the likelihood that agents must repeatedly traverse the same locations—such as narrow corridors or shared entry points—while pursuing new coordination goals. For example, a map layout that requires one agent to backtrack through a previously visited corridor to unlock a second door creates high revisit pressure. We use the parameter $\beta \in [0, 1]$ to characterize geometric pressure in each environment, where higher values indicate more constrained layouts that enforce frequent backtracking.

3.2. Theoretical Hypothesis

We hypothesize that coordination failure arises when agents must revisit earlier checkpoints under sparse rewards, but lifelong novelty discourages such revisits. This decay effect is exacerbated as the number of required coordination points (denoted L) increases, and when the geometry of the environment imposes frequent backtracking (denoted β). We analyze these effects in a communication-free setting, allowing us to isolate the impact of novelty strategy alone.



(a) Grid Environment (b) Grid Environment Results

Figure 1: Overview of the environments and corresponding results. In (b) The solid line represents **Summation of local novelty** and the dotted line represents **MACE**

3.3. Gridworld Design

We design multiple two-agent gridworld tasks (Figure 1a). The tasks in this two-agent environment require coordination via switches and doors to get both agents into a target room. We define a single target reaching or a single switch-door mechanism as one coordination checkpoint. This allows us to precisely test the effect of coordination complexity (L) and geometric revisit pressure (β).

- **`Maze`:** This map only require one coordination event, two agents arrive at the target room. We classify it as having low coordination complexity ($L = 1$). Therefore, this layout does not require agent to revisit to hit the second checkpoint.
- **`Pass`** , **`Pass2`** and **`Pass3`**: These maps contain a sequence of two switch-door mechanisms, making them tasks with higher coordination complexity ($L = 2$). Their layouts are designed to create a low to high level of geometric pressure. `Pass3`, for instance, forces one agent to backtrack over the starting area to reach the second door, inducing a high revisit-rate exponent ($\beta \approx 1$).
- **`SecretRoom`:** Another map with a multi-step coordination sequence ($L = 2$) used to further validate our findings on complex tasks.
- **`DoubleRoom`:** This map is a extended version of **`Pass3`**, which contains more coordination sequence ($L > 2$) used to further validate our findings on complex tasks.

3.4. Approach

We hypothesize that the effectiveness of intrinsic motivation in decentralized MARL depends on two structural properties: the task's coordination complexity (L) and geometric revisit pressure (β). To explore this, we designed tasks that vary these two factors. To test in practical settings, our experiment included communication. We evaluate three main novelty schemes—*Lifelong only*, *Episodic only*, and *Combined* (our hybrid approach)—under two communication settings:

- **Summation of local novelty:** $r^i_{\text{int}} = u^i_t + \sum_{j \neq i} u^j_t$
- **MACE [14]:** $r^i_{\text{int}} = u^i_t + \sum_{j \neq i} u^j_t + \lambda \sum_{j \neq i} v_{i,j}$

Here u^i_t is the intrinsic motivation agent i get and u^i_t is the intrinsic motivation it received from agent j at time t. Compared to the summation of local novelty, MACE [14] has one more component $v_{i,j}$ to measure the influence of agent i 's action on the other agent j 's novelty. For *Lifelong* bonus, we used $u^i = \sqrt{N_k(s^i)}$ to measure the novelty of a state s^i for agent i, while N_k is the accumulated visit counts on this state till k episodes. For *Episodic* bonus, we used $u^i = \mathbb{I}[N_e(s^i) = 1]$ to measure the novelty of a state within one episode, where $N_e(s^i)$ is the visit counts of the state within current episode. To combine the global and episodic bonuses, we use multiplication, as it is used in [20, 15], therefore, the combined intrinsic reward is: $u^i = \frac{1}{\sqrt{N_k(s^i)}} \cdot \mathbb{I}[N_e(s^i) = 1)]$

4. Initial Experimental Results

The GridWorld experiments provide the most direct validation of our theory by systematically varying the coordination complexity (L) and geometric revisit pressure (β). The results in Figure 1b show a clear phase transition consistent with our predictions. In the simple `Maze` task ($L = 1$), the lifelong-only strategy performs well and the performance is very close to episodic-only and combined one, as expected. In `Pass` task ($L = 2$) map, which increase the sequential complexity, the performance of the lifelong-only agent is still well, only slightly worse than the combined novelty one. However, in the `Pass2` and `Pass3` maps, which increase the sequential complexity ($L = 2$) and use layouts that force backtracking (high β), the performance of the lifelong-only agent degrades significantly. In stark contrast, the *Combined* approach, featuring an episodic bonus, maintains its high performance across all these challenging maps. This demonstrates that an episodic component becomes essential as task and geometric complexity rise. These trends align well with our theoretical predictions, confirming that the interaction between L and β is critical to the success of different intrinsic motivation schemes.

5. Conclusion

Our work identifies coordination de-synchronisation as a critical failure mode in decentralised multi-agent exploration. We show that exploration strategy effectiveness depends on two key factors: the task's logical complexity (L) and the environment's geometric revisit pressure (β). We hope this work sparks further investigation into coordination-aware exploration in decentralized settings. In particular, we invite the community to consider new forms of episodic or structure-aware intrinsic motivation, and to develop theoretical tools that capture the interplay between coordination structure and exploration dynamics. Addressing these challenges is crucial for scaling MARL to real-world applications that demand reliable and sustained coordination.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Xu, F. Zhong, Y. Wang, Learning multi-agent coordination for enhancing target coverage in directional sensor networks, Advances in Neural Information Processing Systems 33 (2020) 10053–10064.

[2] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, L. Hanzo, Multi-agent deep reinforcement learning-based trajectory planning for multi-uav assisted mobile edge computing, IEEE Transactions on Cognitive Communications and Networking 7 (2020) 73–84.

[3] M. Zhou, X. Ma, Y. Li, A novel multi-objective routing scheme based on cooperative multi-agent reinforcement learning for metaverse services in fixed 6g, in: WOCN, 2023. URL: https://dblp.org/rec/conf/wocc/ZhouML23, dOI: 10.1109/WOCC52294.2023.00029.

[4] C. S. De Witt, T. Gupta, D. Makoviichuk, V. Makoviychuk, P. H. Torr, M. Sun, S. Whiteson, Is independent learning all you need in the starcraft multi-agent challenge?, arXiv preprint arXiv:2011.09533 (2020).

[5] J. Jiang, Z. Lu, I2q: A fully decentralized q-learning algorithm, Advances in Neural Information Processing Systems 35 (2022) 20469–20481.

[6] T. Zhu, Y. Jin, J. Houssineau, G. Montana, Mitigating relative over-generalization in multi-agent reinforcement learning, arXiv preprint arXiv:2411.11099 (2024).

[7] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, R. Munos, Unifying count-based exploration and intrinsic motivation, Advances in neural information processing systems 29 (2016).

[8] D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, Curiosity-driven exploration by self-supervised prediction, in: International conference on machine learning, PMLR, 2017, pp. 2778–2787.

[9] Y. Burda, H. Edwards, A. Storkey, O. Klimov, Exploration by random network distillation, arXiv preprint arXiv:1810.12894 (2018).

[10] S. Iqbal, F. Sha, Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning, arXiv preprint arXiv:1905.12127 (2019).

[11] T. Wang, J. Wang, Y. Wu, C. Zhang, Influence-based multi-agent exploration, arXiv preprint arXiv:1910.05512 (2019).

[12] L. Zheng, J. Chen, J. Wang, J. He, Y. Hu, Y. Chen, C. Fan, Y. Gao, C. Zhang, Episodic multi-agent reinforcement learning with curiosity-driven exploration, Advances in Neural Information Processing Systems 34 (2021) 3757–3769.

[13] B. Liu, Z. Pu, Y. Pan, J. Yi, Y. Liang, D. Zhang, Lazy agents: a new perspective on solving sparse reward problem in multi-agent reinforcement learning, in: International Conference on Machine Learning, PMLR, 2023, pp. 21937–21950.

[14] H. Jiang, Z. Ding, Z. Lu, Settling decentralized multi-agent coordinated exploration by novelty sharing, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 17444–17452.

[15] M. Toquebiau, N. Bredeche, F. Benamar, J.-Y. Jun, Joint intrinsic motivation for coordinated exploration in multi-agent deep reinforcement learning, arXiv preprint arXiv:2402.03972 (2024).

[16] T. Zhang, H. Xu, X. Wang, Y. Wu, K. Keutzer, J. E. Gonzalez, Y. Tian, Noveld: A simple yet effective exploration criterion, Advances in Neural Information Processing Systems 34 (2021) 25217–25230.

[17] M. Henaff, R. Raileanu, M. Jiang, T. Rocktäschel, Exploration via elliptical episodic bonuses, Advances in Neural Information Processing Systems 35 (2022) 37631–37646.

[18] T. Lin, J. Huh, C. Stauffer, S. N. Lim, P. Isola, Learning to ground multi-agent communication with autoencoders, Advances in Neural Information Processing Systems 34 (2021) 15230–15242.

[19] Y. L. Lo, B. Sengupta, J. Foerster, M. Noukhovitch, Learning multi-agent communication with contrastive learning, arXiv preprint arXiv:2307.01403 (2023).

[20] M. Henaff, M. Jiang, R. Raileanu, A study of global and episodic bonuses for exploration in contextual mdps, in: International Conference on Machine Learning, PMLR, 2023, pp. 12972–12999.