# Emergent Reciprocity Through Temporal Credit Assignment in Reinforcement Learning Agents

**Le Thuy Duong Nguyen** (le.thuy.nguyen@mail.mcgill.ca)
Honours Cognitive Science Program of McGill University, Mila - Quebec Artificial Intelligence Institute
Montreal, QC, Canada

**Dane Malenfant**
School of Computer Science of McGill University, Mila - Quebec Artificial Intelligence Institute
Montreal, QC, Canada

**Blake Aaron Richards**
School of Computer Science, Department of Neurology & Neurosurgery, and Montreal Neurological Institute-Hospital of McGill University, Mila - Quebec Artificial Intelligence Institute
Montreal, QC, Canada
CIFAR Learning in Machines and Brains Program
Toronto, ON, Canada

## Abstract

**Humans and animals have complex social and cultural systems that can span large distances and times. In North America, Plains Indigenous nations practiced reciprocity through *Manitokan*, effigies placed at fixed locations, where leaving surplus goods was seen as cooperative acts of resource sharing or caching. In multi-agent reinforcement learning (MARL), reciprocity has largely been defined as an emergent property through tit-for-tat policies or reputation scores that establish social norms. These perspectives fail to consider the temporal structure of rewards or the criticality of certain actions necessary for success. We present a novel MARL environment to investigate the emergence of reciprocal prosocial behaviours in reinforcement learning agents. Baseline experiments show that agents consistently converged to suboptimal policies favoring individual resource maximization, despite the potential for improved collective outcomes. These findings highlight a critical gap in existing MARL methods, suggesting the need for new algorithms capable of supporting temporal credit assignment in artificial agents.**



**Keywords:** social intelligence; multi-agent reinforcement learning; credit assignment; episodic memory


## Introduction & Methods

Humans and animals are capable of exhibiting complex collective behaviours across long spatial and temporal scales. One such behaviour is reciprocity, or the act of giving without an explicit arrangement of immediate or future returns, but an expectation that others would do the same. This appears in both human cultural practices and animal social systems. Reciprocity is evident in Plains Indigenous traditions through *Manitokan* (Barkwell, n.d.), isolated effigies provisioned with resources such as food, medicine, and tools to benefit passing travelers. Leaving unneeded goods at these can be perceived as cooperative actions; a reciprocal sharing of resources or caching of items (Clayton & Dickinson, 1998). Fixed locations are episodic memory-efficient and reduce uncertainty when traveling. Current MARL perspectives fail to consider the temporal structure of rewards or the criticality of certain actions necessary for success (Sutton, 1984).

### Environment Design

To investigate the emergence of reciprocal, prosocial behaviours in artificial agents, we introduce a novel MARL task under development based in MiniGrid (Chevalier-Boisvert et al., 2023), in which decentralized agents navigate a discrete world partitioned into heterogeneous quadrants. Each quadrant contains distinct resources (four berry types) and salient visual cues. At the center, tree objects represent the culturally-inspired *Manitokan*, a focal area for resource sharing. Adjustable parameters allow systematic exploration of various configurations, including resource availability, density, and grid size (Figure 1). Agents operate in a discrete action space that includes turning left and right, moving forward, and picking up or dropping berries.

### Delayed Exponential Reward

Each agent $i$ receives an individual reward at the end of each episode based on the number of unique berry types collected, $n_i$. The reward is computed as an exponentially increasing function of $n_i$, and a constant step penalty, λ, is applied at every timestep. The optimal policy is defined as one that maximizes resource diversity while minimizing extraneous movement:

$$R_i = \begin{cases} 2^{n_i - 1} - \lambda, & \text{if } n_i > 0 \\ -\lambda, & \text{if } n_i = 0 \end{cases} \quad \forall i \tag{1}$$

Since the rewards are given at the end of an episode, agents do not receive a dense signal for facilitating another agent's success to which they need to assign credit. This delayed exponential reward promotes emergent strategies that favor memory-efficient and diverse resource collection, offering a testbed for evaluating social and cooperative behaviour under decentralized partial observability.

## Results

Baseline experiments with proximal policy optimization (PPO) (Schulman et al., 2017; Yu et al., 2022) indicate that agents primarily exploit their local quadrant, resulting in suboptimal individual policies with minimal

exploration despite the potential for higher collective rewards. As shown in Figure 2, Policy loss metrics consistently decreased throughout the training, indicating improving stability in learned policies. Value network losses diminished steadily, reflecting improved accuracy in estimating future rewards. Despite these training improvements, agents converged to stable yet suboptimal policies characterized by selfish resource collection strategies rather than engaging in reciprocal interactions to enhance collective reward outcomes (Figure 3).

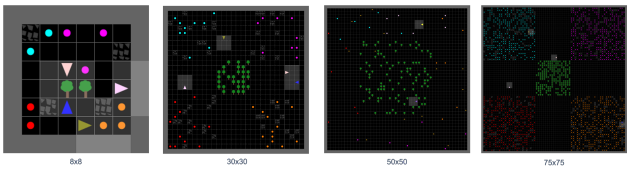


Figure 1: MARL environments constructed with adjustable parameters for grid size, resource density, and spatial structure. Each environment consists of four quadrants with distinct berry objects and central tree objects placed at their intersection. Agent and object positions are randomly initialized within quadrants at the start of each episode.

We trained four agents in the 30x30 environment for 7,000 episodes using the delayed exponential reward scheme. The average number of steps during which berries were held increased rapidly during early training and plateaued after ~3,000 episodes, while the drop rate declined steadily to near zero. This behaviour indicates a shift from exploratory to conservative strategies where agents retain collected resources without sharing them at the central *Manitokan*. The observed suboptimal equilibrium highlights an interesting gap within current decentralized MARL frameworks.

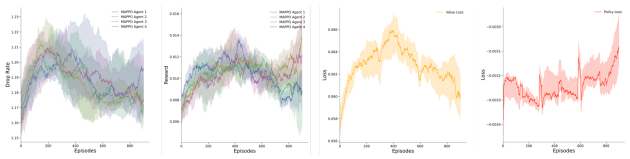

Figure 2: Training dynamics of MAPPO agents. Initial cooperative behaviours in early training marked by increased berry drop rates and growing reward returns across all four agents. Shaded regions denote interquartile ranges across 10 seeds.

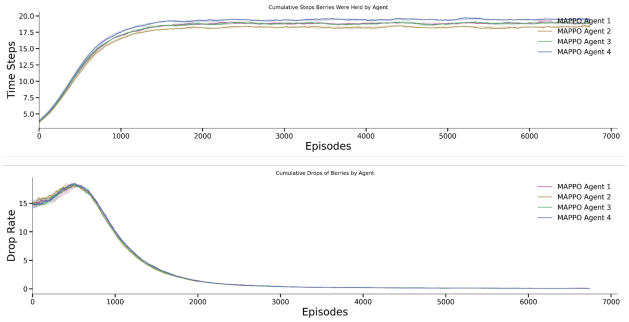


Figure 3: Behavioural convergence of MAPPO agents. Across training, agents increase hoarding behaviour (top) and reduce cooperative resource dropping (bottom).

## Conclusion & Future Work

Existing MARL approaches fail to develop cooperative trading under natural constraints despite the potential for increased collective rewards. Our environment and results contribute novel evaluation tools and establish benchmarks crucial to overcome this gap and advance our understanding of social intelligence in both biological and artificial systems. Future work may develop a human-controllable version of the environment to facilitate parallel psychological studies that compare human social strategies to those emerging in artificial agents.

## Acknowledgments

BAR received support from NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031) and CIFAR (Canada AI Chair; Learning in Machine and Brains Fellowship). DM received support from an NSERC CGSM and a Rathlyn Fellowship from the Indigenous Studies Department of McGill. This research was enabled in part by support provided by (Calcul Québec) (https://www.calculquebec.ca/en/) and the Digital Research Alliance of Canada (https://alliancecan.ca/en). The authors acknowledge the material support of NVIDIA in the form of computational resources.

## References


Barkwell, L. (n.d.). *Manitokanac* [PDF]. Gabriel Dumont Institute of Native Studies and Applied Research Virtual Métis Museum. Retrieved April 7, 2025, from https://www.metismuseum.ca/resource.php/148154

Chevalier-Boisvert, M., Dai, B., Towers, M., Perez-Vicente, R., Willems, L., Lahlou, S., ... & Terry, J. (2023). Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, *36*, 73383-73394.

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

Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, *35*, 24611-24624.