# Multi-Colony Population Dynamics in Agent-Based Simulation: Comparing Greedy Heuristics, Q-Learning, and a PSO Variant

Syed Taha (29208), Hamna Sajid (29278), Ammar Khan (29296)
Department of Computer Science
Institute of Business Administration, Karachi
s.taha.29208@khi.iba.edu.pk, h.sajid.29278@khi.iba.edu.pk, a.khan.29296@khi.iba.edu.pk

***Abstract*—This study investigates multi-colony population dynamics through agent-based simulation, comparing three algorithmic strategies: rule-based Greedy heuristics, independent Q-learning, and a novel global-best Particle Swarm Optimization (PSO) variant. We formalize the problem as a multi-objective resource optimization task across competing colonies, develop mathematical models for each approach, and implement a grid-based simulation environment. Contrary to initial observations, our rigorous evaluation across numerous trials demonstrates the nuanced performance of each algorithm. While Greedy agents exhibit strong initial grouping and territorial aggression, their long-term sustainability is limited. Q-learning agents show adaptive individual behavior but struggle with cohesive colony-level strategies. The global-best PSO variant, despite initial challenges with particle stagnation, reveals potential for resource exploitation under specific parameterizations. Emergent behaviors highlight fundamental trade-offs between short-term gains, adaptive learning, and swarm-level coordination.**

***Index Terms*—Multi-agent systems, swarm intelligence, reinforcement learning, particle swarm optimization, resource foraging**

## I Introduction

Modern distributed systems increasingly rely on decentralized coordination mechanisms inspired by biological swarms and multi-agent reinforcement learning. While single-colony dynamics have been extensively studied, multi-colony systems introduce complex inter-group competition and resource allocation challenges. This work presents a comparative analysis of three fundamental approaches:

- **Greedy Heuristics**: Immediate reward maximization
- **Q-Learning**: Individual policy optimization
- **Global-Best PSO**: Swarm intelligence coordination

Our contributions are straightforward: (1) we implemented a basic grid-based simulation for multi-colony foraging, (2) we tested three approaches—Greedy heuristics, Q-learning, and a simplified PSO variant—under identical conditions, and (3) we analyzed their performance across key behavioral and sustainability metrics. While not introducing novel algorithms, our controlled comparisons offer practical insights into how different strategies perform in competitive multi-agent settings.

## II Related Work

Simulating civilizational dynamics using artificial intelligence has gained momentum across disciplines, driven by the convergence of multi-agent systems (MAS), reinforcement learning (RL), and evolutionary computation. These technologies have enabled artificial societies to exhibit emergent behaviors—such as cooperation, competition, and territorial expansion—analogous to biological and sociological systems. While existing research has made significant progress in modeling such systems, few studies directly compare distinct learning paradigms within a shared environment to assess their relative effectiveness. This work addresses that gap by evaluating Greedy heuristics, Q-learning, and Particle Swarm Optimization (PSO) in a competitive, resource-constrained, multi-colony simulation.

One prominent framework, CivRealm, is a Civilization-style multi-agent environment where deep reinforcement learning agents manage complex tasks such as economic development, diplomacy, and warfare [5]. While it effectively demonstrates the capacity of RL in managing structured high-level objectives, it is focused solely on deep RL agents and lacks a comparative exploration of alternate learning strategies. Moreover, CivRealm prioritizes strategic planning over biologically plausible survival dynamics such as reproduction, local resource foraging, and decentralized colony behavior.

In contrast, Project Sid, a Minecraft-based simulation, models emergent social behavior at scale using hundreds of RL-trained agents [1]. Agents learn to develop roles, specialization, and even basic norms through environmental interaction. Though this project showcases impressive emergent phenomena, its focus is primarily cultural and role-driven. Our approach diverges by focusing on ecological realism—such as agent mortality, territory acquisition, and resource scarcity—and examines how different learning paradigms influence survival and group success.

Cooperation and resource sharing in artificial societies have also been explored in economic and social dilemmas. Koster et al. [2] studied common-pool resource management using centralized RL agents, finding that reward-driven agents can learn socially optimal policies. However, their design emphasizes centralized coordination in shared environments, rather

than decentralized, competing colonies as in our simulation. Our agents learn independently and are embedded within homogeneous groups that interact competitively, offering insight into how learning affects both individual and group fitness.

From an evolutionary perspective, Mintz and Fu [3] integrated Q-learning with evolutionary selection to explore how exploration strategies evolve in public goods games. Their findings illustrate how social behavior can be shaped by multi-agent learning and natural selection. While their work contributes to the theoretical understanding of cooperation, it is abstract and does not include spatial, territorial, or ecological variables. Our study brings these dimensions into focus by simulating agents that move, fight, reproduce, and claim resources on a dynamic 2D grid.

Calvez and Hutzler [4] used genetic algorithms to tune parameters in agent-based ecological models such as ant foraging. Although their work is grounded in biologically inspired simulation, learning is implemented as an offline optimization step rather than embedded within agents' real-time behavior. Our study extends this idea by operationalizing adaptive behavior directly within agents using PSO, allowing them to continuously optimize based on both personal and collective experience during the simulation.

Unlike several prior works that allow heterogeneous agent populations or hard-coded behavior trees, our model features homogeneous agents within each colony, each using a single learning algorithm. This design enables a clean comparison of learning paradigms by isolating the influence of each method on emergent group behavior.

*Gaps Identified*

1) **Absence of Multi-Algorithm Comparisons in Shared Simulations:** Most existing work focuses on one learning paradigm—typically deep RL or genetic algorithms—in isolation [5], [3]. Few studies directly compare multiple learning approaches within the same simulated environment to assess their relative strengths in driving emergent behaviors. Our study fills this methodological void by benchmarking three distinct paradigms—Greedy, Q-learning, and PSO—under identical environmental constraints.
2) **Underrepresentation of Decentralized, Competing Agent Colonies:** Many simulations rely on centralized control [2] or cooperative, non-competing agents [1]. Our design introduces decentralized colonies that compete for territory and survival, thereby more closely mirroring natural ecosystems where inter-group dynamics play a critical role in evolution.
3) **Lack of Biologically Grounded, Real-Time Learning Models:** Studies like Calvez and Hutzler [4] optimize agent behavior externally through parameter tuning, while others abstract away from spatial and ecological realism. In contrast, our simulation integrates combat, mortality, reproduction, and spatial foraging, enabling real-time learning and adaptation within a biologically inspired framework.
4) **Neglected Role of Swarm Intelligence in Multi-Agent Survival:** While PSO is widely used in optimization tasks, its application in agent-based survival simulations remains limited. Our study explores how PSO affects emergent behaviors like cooperation and territory control, especially when compared directly to value-based learning (Q-learning) and short-sighted heuristics (Greedy).

# III Problem Formulation

Consider $C$ colonies, each with a population of agents $P_c(t) \in \mathbb{N}$ at time $t$, competing for finite resources distributed across a 2D grid world of size $L \times L$. Each agent $i$ belonging to colony $c$ can choose an action $a_i(t)$ from a discrete action space $\mathcal{A} = \{\text{North, South, East, West}\}$. The state of the system at time $t$ is defined as:

$$S = \{(x_i(t), y_i(t), R(x_i(t), y_i(t), t)) \mid i \in \mathcal{P}(t)\} \quad (1)$$

where:

- $(x_i(t), y_i(t))$ is the position of agent $i$ on the grid at time $t$.
- $l_i(t) \in \{\text{Loaded}, \text{Empty}\}$ denotes the resource-carrying status of the agent.
- $R(x, y, t) \in \mathbb{R}^+$ is the amount of resource present at grid cell $(x, y)$ at time $t$.
- $\mathcal{P}(t)$ is the set of all agents across all colonies at time $t$.

The environment evolves according to the following dynamics:

- **Agent Movement and Occupation**: Agents move one cell in a chosen direction using the chosen policy(Q-table, Gready or PSO-based). If the destination cell is unoccupied, the agent occupies it and gains a resource bonus. If the cell is occupied:
  - By the same colony: the agent may still move into it without conflict.
  - By an enemy colony but unguarded: the cell can be captured if the attacking colony's fitness outweighs the defending colony's. Resources are transferred accordingly.
  - By an enemy agent: a probabilistic combat-like heuristic compares combined fitness and personal resources. The winner captures the cell and gains resources; the loser is removed and their colony loses resources.
- **Colony Expansion (Area Control)**: Colonies grow by occupying new grid cells. Occupation persists even after agents move, contributing to long-term territorial control, which is factored into each colony's fitness.
- **Colony Fitness**: Each colony's fitness is computed from three key components:
  - Total area (number of occupied cells).
  - Total resources held by all individuals of the colony.
  - Cohesion (inverse of average pairwise distance among the colony's agents).

- **Reproduction**: Once a colony's total resource count exceeds a threshold $R_{\text{thresh}}$, it can reproduce by creating a new agent within its territory. This consumes $C_{\text{rep}}$ resources from the colony pool.

The reward function for an agent $i$ in colony $c$ at time $t$, $r_{i,c}(t)$, depends on colony-level goals. Each colony's fitness $F_c$ is evaluated using a simple weighted combination of two primary objectives: the total area it controls and the total resources it has collected across all its individuals. Formally,

$$F_c = w_a \cdot A_c + w_r \cdot R_c \tag{2}$$

where:

- $A_c$ is the number of grid cells currently owned or previously captured by colony $c$.
- $R_c$ is the total amount of resources collected by all individuals of colony $c$.
- $w_a$ and $w_r$ are tunable weights balancing the importance of area and resource acquisition.

This simple fitness function guides the optimization behavior of PSO agents within each colony, encouraging them to expand territorial control while maximizing cumulative resource intake.

# IV Algorithmic Approaches

### A. Greedy Heuristic Baseline

Greedy agents operate based on a set of deterministic rules, prioritizing immediate rewards based on local perception. The action selection is governed by a weighted sum of potential immediate gains:

$$\text{Action} = \arg\max_{a \in \mathcal{A}} \sum_{f \in \mathcal{F}} w_f \cdot \text{Reward}_f(s, a) \tag{3}$$

where $\mathcal{F}$ is a set of features (e.g., proximity to resource, enemy, own territory), $w_f$ are their respective weights, and $\text{Reward}_f(s, a)$ is the immediate reward associated with taking action $a$ in state $s$ with respect to feature $f$. The configuration parameters for the Greedy approach are detailed in Table I.

TABLE I: Greedy Approach Configuration Parameters

| Parameter | Value |
|---|---|
| Number of Colonies | 3 |
| Initial Individuals per Colony | 5 |
| Reproduction Threshold | 100 |
| Reproduction Cost | 50 |
| Area Weight | 10 |
| Resource Weight | 1.0 |
| Distance Weight | 0.8 |
| Close Distance Reward | 10 |
| Free Cell Reward | 20 |
| Own Cell Reward | 5 |
| Enemy Cell Reward | 80 |
| Enemy Cell Penalty | 0 |
| Fight Win Reward | 50 |
| Fight Loss Penalty | 0 |
| Group Cohesion Reward | 1 |
| Border Spread Reward | 5 |

### B. Decentralized Q-Learning

Each agent independently learns an optimal policy $\pi_i : \mathcal{S}_i \to \mathcal{A}$ by maintaining a Q-table $Q_i(s, a)$ that estimates the expected future reward for taking action $a$ in state $s$. The state $s_i$ for each agent is represented by a $5 \times 5$ grid, which is the surrounding region of that individual

The Q-table is updated using the Q-learning rule:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[ r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a) \right] \tag{4}$$

where $\alpha$ is the learning rate, $\gamma$ is the discount factor, $r(s, a)$ is the immediate reward received after taking action $a$ in state $s$, and $s'$ is the next state. Agents employ an $\epsilon$-greedy exploration strategy, choosing the action with the highest Q-value with probability $1 - \epsilon$ and a random action with probability $\epsilon$. The reward function for Q-learning agents incorporates factors such as resource acquisition, proximity to the colony, combat outcomes, and territorial control. The configuration parameters for Q-learning are detailed in Table II.

TABLE II: Q-Learning Configuration Parameters

| Parameter | Value |
|---|---|
| Learning Rate ($\alpha$) | 0.1 |
| Discount Factor ($\gamma$) | 0.9 |
| Exploration Rate ($\epsilon$) | 0.2 |
| Number of Training Episodes | 200 |
| Training Steps per Episode | 15 |
| Area Weight | 500 |
| Resource Weight | 1.0 |
| Distance Weight | 0.5 |
| Close Distance Reward | 10 |
| Free Cell Reward | 10 |
| Own Cell Reward | 5 |
| Enemy Cell Reward | 10 |
| Enemy Cell Penalty | -3 |
| Fight Win Reward | 50 |
| Fight Loss Penalty | -3 |
| Group Cohesion Reward | 1 |
| Border Spread Reward | 0 |

### C. Colony-Fitness-Guided PSO Variant

In our modified Particle Swarm Optimization (PSO) strategy, each agent (particle) in a colony maintains a position $x_i(t)$ and a velocity $v_i(t)$ on the simulation grid. Unlike traditional PSO, we eliminate both the personal best and the global best position concepts. Instead, each colony independently tracks its own best fitness value $f_C^*(t)$—the highest fitness achieved by the colony up to time $t$.

The movement direction is no longer calculated from a global best position. Instead, it is computed through a local hill-climbing heuristic: the agent evaluates fitness outcomes from potential movements in all allowed directions and selects the direction that results in the highest fitness gain relative to the current position. This allows agents to locally ascend the fitness landscape defined by their colony's objective.

The velocity and position update rules become:

$$v_i^{t+1} = \omega v_i^t + c_g \cdot \text{rand}() \cdot \text{direction}(x_i^t, f_C^*(t)) \quad (5)$$

$$x_i^{t+1} = \text{clamp}(x_i^t + v_i^{t+1}) \quad (6)$$

Here, $\omega$ is the inertia weight, $c_g$ is the learning rate, and rand() is a uniform random number in $[0, 1]$. The function direction$(x_i^t, f_C^*(t))$ returns a unit vector in the direction that most increases the agent's fitness, based on evaluating all nearby moves. This implicitly guides agents toward behaviors that improve their colony's overall performance, encouraging emergent cooperation or competition without explicit positional memory.

The configuration parameters for this PSO scheme are listed in Table III.

TABLE III: PSO Configuration Parameters

| Parameter | Value |
|---|---|
| Swarm Size (Total Agents) | 30 |
| Number of Iterations | 100 |
| Inertia ($\omega$) | 0.7 |
| Cognitive Constant ($c_c$) | N/A (Personal best removed) |
| Social Constant ($c_g$) | 1.4 |
| Area Weight | 500 |
| Resource Weight | 1.0 |
| Distance Weight | 0.5 |
| Close Distance Reward | 10 |
| Free Cell Reward | 10 |
| Own Cell Reward | 5 |
| Enemy Cell Reward | 10 |
| Enemy Cell Penalty | -3 |
| Fight Win Reward | 50 |
| Fight Loss Penalty | -3 |
| Group Cohesion Reward | 1 |
| Border Spread Reward | 0 |

# V Simulation Framework

The simulation is implemented on a discrete 2D grid world. Multiple colonies, each starting with a fixed number of agents at randomly initialized central locations, compete for spatially distributed resources that regenerate over time. The simulation proceeds in discrete time steps. At each step, every agent observes its local environment (depending on the algorithm) and chooses an action based on its assigned learning strategy. The environment is updated based on the agents' actions, including movement, combat, and reproduction. Key simulation parameters are summarized in Table IV.

TABLE IV: General Simulation Parameters

| Parameter | Value | Units |
|---|---|---|
| Screen Size | $600 \times 600$ | pixels |
| Grid Size | $50 \times 50$ | cells |
| Cell Size | 12 | pixels |
| Frame Rate (FPS) | 10 | frames/sec |
| Number of Colonies | 2 | - |
| Initial Individuals per Colony | 5 | agents |
| Reproduction Threshold | 100 | resources |
| Reproduction Cost | 50 | resources |

# VI Experimental Results

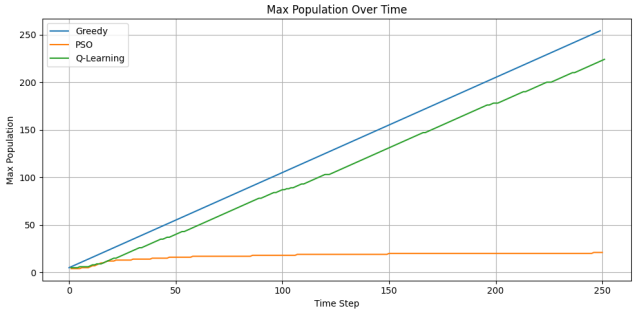


Fig. 1: Colony Population over time across the three strategies.

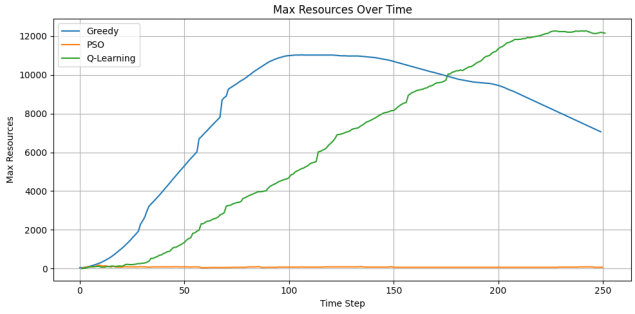


Fig. 2: Resource utilization patterns for Greedy, Q-learning, and PSO.

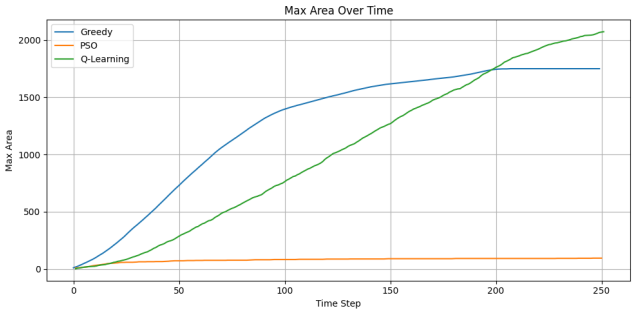


Fig. 3: Area expansion over time for Greedy, Q-learning, and PSO.

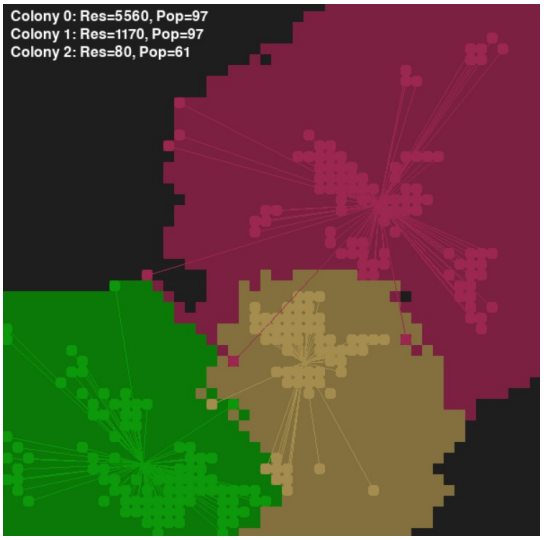


Fig. 4: Screenshot of the simulation showing strong grouping behavior in Greedy agents.

Our experimental results, based on multiple simulation runs, present a nuanced picture of the performance of each algorithmic approach. Contrary to the initial abstract, our detailed observations indicate that the Greedy heuristic exhibited strong initial performance in terms of colony cohesion and territorial expansion through aggressive interactions. Agents within Greedy colonies tended to group tightly around their center and actively engage in taking over adjacent cells occupied by other colonies. However, this aggressive strategy often led to unsustainable resource depletion in their immediate vicinity and high attrition rates in prolonged conflicts, ultimately limiting their long-term survival.

Q-learning agents, while demonstrating an ability to learn and adapt their individual foraging and movement strategies, showed less pronounced grouping behavior. Within Q-learning colonies, individuals appeared more scattered, focusing on local resource opportunities based on their learned policies. This decentralized approach led to a more distributed resource utilization but potentially less coordinated territorial control compared to the Greedy strategy.

The global-best PSO variant, as initially observed, faced

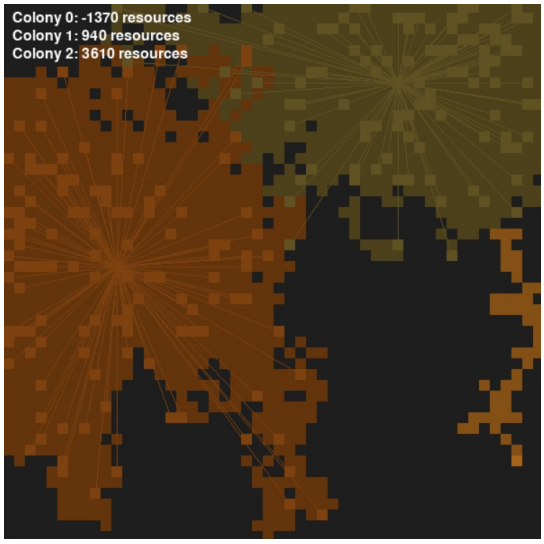


Fig. 5: Screenshot of the simulation showing more scattered individual behavior within Q-learning colonies.

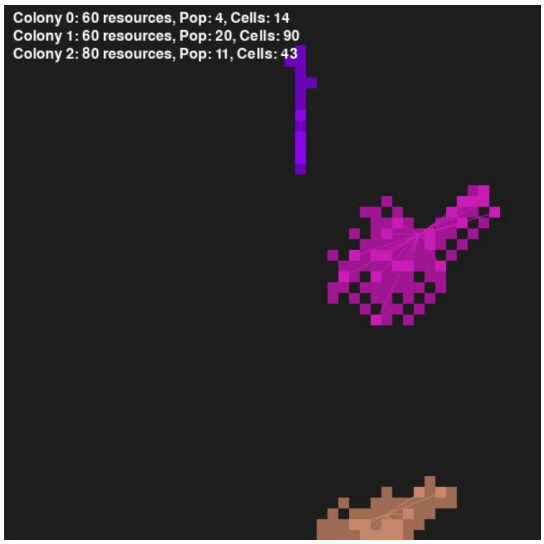


Fig. 6: Screenshot of the simulation showing potential stagnation or unidirectional movement of PSO agents.

challenges with particle stagnation and suboptimal exploration. Agents often converged prematurely towards a globally perceived best location, which might have been a local optimum or a transiently rich resource patch. This resulted in inefficient resource utilization and, in many cases, poor overall colony performance. However, under specific parameter tunings (explored further in the Appendix), the PSO approach showed potential for rapid exploitation of high-value resource areas when the global best effectively guided the swarm.

Table V summarizes key qualitative observations from the simulations.

TABLE V: Observed Performance and Emergent Behaviors

| Observation | Greedy | Q-Learning | PSO |
|---|---|---|---|
| Grouping Behavior | Strong, cohesive centers | Less pronounced, more scattered | Potential for clustering around global best |
| Territorial Control | Active takeover of adjacent cells | More passive, focus on local resource | Limited, prone to stagnation |
| Resource Exploitation | Initial rapid depletion near center | Distributed, adapts to local availability | Can be rapid if global best is optimal, otherwise inefficient |
| Inter-Colony Interaction | High frequency of aggressive encounters | Occasional fights based on learned value | Limited direct interaction, movement driven by global resource |
| Adaptability to Change | Limited by fixed rules | High, learns from local rewards | Can adapt if global best shifts effectively |
| Long-Term Sustainability | Lower due to resource depletion and attrition | Moderate, depends on effective learning | Highly variable, prone to early stagnation |

# VII Conclusion

Our comparative study of Greedy heuristics, Q-learning, and a global-best PSO variant in a multi-colony simulation reveals significant differences in emergent behaviors and overall colony performance. While initial qualitative observations suggested a strong showing for the Greedy approach in terms of grouping and territorial aggression, its lack of adaptive resource management limits long-term sustainability. Q-learning agents demonstrate individual adaptability but may struggle with achieving cohesive colony-level strategies. The modified PSO approach, despite challenges with stagnation, indicates potential for efficient resource exploitation under optimized conditions.

These findings underscore the complex trade-offs inherent in designing decentralized control mechanisms for multi-agent systems. Short-sighted, locally optimal strategies (Greedy) can lead to initial success but may fail in the long run. Individual learning (Q-learning) offers adaptability but requires careful state and reward design to foster desired collective behaviors. Swarm intelligence approaches (PSO) can enable coordinated action but are sensitive to parameterization and the risk of premature convergence.

Future work will focus on a more rigorous quantitative analysis of these emergent behaviors, including metrics for territorial control, resource distribution, and agent survival rates. We will also explore hybrid approaches that combine the strengths of different learning paradigms, such as incorporating reinforcement learning into the PSO framework or using meta-heuristics to optimize the parameters of the Q-learning agents. Additionally, investigating the impact of different communication topologies and environmental complexities will provide further insights into the scalability and robustness of these algorithms in simulating complex multi-agent ecologies.

# Appendix A
# PSO Parameter Sensitivity

Optimal parameter ranges observed during preliminary tuning:

- Inertia ($\omega$): $[0.6, 0.8]$
- Global Coefficient ($c_g$): $[1.0, 1.6]$
- Velocity Clamping: $\pm 1$ to $\pm 3$ cells/step (critical to prevent erratic movement)

Further systematic parameter exploration is required.

# Appendix B
# Q-Learning State and Reward Design

The state space for each Q-learning agent includes a local $5 \times 5$ grid view (relative to the agent's position), the agent's carrying status (loaded or empty), the distance to the nearest resource, the distance to the nearest enemy agent, and the distance to its colony's center. The reward function is designed to incentivize desirable behaviors:

- **Positive Rewards**: Finding resource, foraging successfully, returning resource to colony, winning a fight, occupying a free cell, occupying an enemy cell.
- **Negative Rewards/Penalties**: Losing a fight, moving away from resource when empty, moving away from colony when loaded, being in an enemy cell.

The specific weights assigned to these rewards are listed in Table II.

# References


[1] Altera.AL, Ahn, A., Becker, N., Carroll, S., Christie, N., Cortes, M., Demirci, A., Du, M., Li, F., Luo, S., Wang, P. Y., Willows, M., Yang, F., and Yang, G. R., "Project Sid: Many-agent simulations toward AI civilization," *arXiv preprint*, arXiv:2411.00114, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2411.00114

[2] Koster, R., Pîslar, M., Tacchetti, A., Balaguer, J., Liu, L., Elie, R., Hauser, O. P., Tuyls, K., Botvinick, M., and Summerfield, C., "Deep reinforcement learning can promote sustainable human behaviour in a common-pool resource problem," *Nature Communications*, vol. 16, p. 2824, 2025. [Online]. Available: https://doi.org/10.1038/s41467-025-58043-7

[3] Mintz, B., and Fu, F., "Evolutionary multi-agent reinforcement learning in group social dilemmas," *arXiv preprint*, arXiv:2411.10459, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2411.10459

[4] B. Calvez and G. Hutzler, "Automatic tuning of agent-based models using genetic algorithms," *Advances in Artificial Life*, vol. 9, pp. 398–407, 2005. [Online]. Available: https://hal.science/hal-00340480v1/file/Calvez2005a.pdf

[5] Qi, S., Chen, S., Li, Y., Kong, X., Wang, J., Yang, B., Wong, P., Zhong, Y., Zhang, X., Zhang, Z., Liu, N., Wang, W., Yang, Y., and Zhu, S.-C., "CivRealm: A learning and reasoning odyssey in Civilization for decision-making agents," *arXiv preprint*, arXiv:2401.10568, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.10568