Project Report Part 2:

ClanQuest: Emergence of Geopolitical Intelligence via Multi-Agent Deep Reinforcement Learning

Team Name: A2S
Arnab Banerjee (B2430040) – Team Lead
Sagnik Biswas (B2430056)
Souvik Deb (B2430061)

Course: DA 346: Deep Reinforcement Learning
Instructor: Tamal Mj
MSc in Big Data Analytics, RKMVERI, Belur

January 5, 2026

Abstract

This project explores the emergence of complex geopolitical behaviors in a multi-agent reinforcement learning (MARL) environment named **ClanQuest**. Built as a custom Gym-compatible simulation, the environment challenges agents to survive in a resource-constrained, seasonally dynamic grid world. Using Proximal Policy Optimization (PPO), we demonstrate how simple individual survival mandates transform into clan-level territorial defense, efficient foraging patterns, and an emergent "Cold War" stalemate. This report provides a comprehensive breakdown of the environment design, MDP specification, algorithm selection, and a deep-dive analysis of the resulting emergent strategies.

Contents

1 Introduction

Reinforcement Learning is often applied to games with fixed boundaries and clear win-state rewards. **ClanQuest** shifts this paradigm toward biological and social intelligence, where the goal is not to "win" in a traditional sense, but to *persist* in a world of limited resources and hostile competitors. By modeling agents with emotional and risk-aware heuristics, we investigate how geopolitical boundaries can emerge purely from the pressure of survival.

2 Environment Design

2.1 A Custom Gymnasium Environment

ClanQuest is implemented as a high-performance Python simulation adhering to the Gymnasium API. The world is a 50x50 discrete grid, supporting multiple concurrent agents. The core design philosophy centers on three pillars:

- **Dynamic Resource Scarcity**: Resources (food) are not static. The world undergoes seasonal cycles (e.g., Spring vs. Winter). During Winter, the resource regeneration rate significantly decreases, forcing agents to have already secured a surplus or established efficient long-range foraging paths.
- **Territorial Governance**: The grid is divided into clan territories. Agents have a "home territory" where metabolic costs are standard ($C = 0.15$). Stepping into a foreign territory triggers a 1.5x metabolic surcharge, modeling the stress and energy cost of evasion or combat in hostile lands.
- **Sovereign Decision Making**: Every agent is an independent PPO policy instance. While they share a clan identity and resource pool, their actions are individual, requiring coordination to arise naturally rather than through hard-coded rules.

2.2 Human Complexity Check

The environment satisfies a high complexity threshold. A human player would struggle to maintain 15 agents simultaneously across 2,000 steps while resources regenerate unpredictably. One must predict seasonal shifts, manage the metabolic drain of traversal, and defend borders—all without explicit communication. The optimal policy is a non-obvious balance of isolationism, foraging efficiency, and defensive alertness.

3 MDP Specification

Formally, the project defines the interaction as a decentralised-observable MDP. Each agent observes a partial segment of the world but acts to optimize a global survival mandate.

3.1 Observation Space

The observation is a 10-dimensional vector $O \in \mathbb{R}^{10}$, normalized for neural network stability:

1. **Spatial Coordination**: $(x/L, y/L)$ coordinates.
2. **Resource Status**: Internal energy / 50.0.
3. **Socio-Emotional State**: Normalized discrete emotion (CALM=0.0 to FEARFUL=1.0).
4. **Growth Trend**: Recent resource delta ΔR.

5. **Global Resource Scent**: A 3D vector $(dx, dy, dist)$ pointing to the nearest known resource field. Crucially, the agents can only "smell" resources in their own territory or within their visual range (Fog of War).

6. **Local Scent**: The intensity of the immediate resource field.

3.2 Action Space

The agents utilize a discrete action space **Discrete(5)**: $A = \{0 : \text{Up}, 1 : \text{Down}, 2 : \text{Left}, 3 : \text{Right}, 4 : \text{Idle}\}$. Movement consumes metabolic energy, while 'Idle' consumes a baseline cost.

3.3 Reward Design: The Core Survival Mandate

Reward design is the most critical component of RL success. We avoided complex reward shaping to prevent "reward hacking" and instead used a hierarchical survival structure:

$$r_t = \mathbb{I}(\text{alive}) \cdot \left(1.0 + 0.5 \times \frac{P_{clan\_alive}}{P_{clan\_total}}\right) \tag{1}$$

If an agent starves ($R \leq 0$), it receives a one-time penalty of -20.0. **Trade-off Analysis**: We initially tried a purely selfish reward ($+1.0$ for self only). This led to agents ignoring their clan members, often blocking each other at resource nodes. By adding the $+0.5 \times \text{ClanRatio}$, agents developed a rudimentary form of spatial awareness, giving way to clan-mates to maximize the collective survival percentage.

3.4 Transition Dynamics

The cell-to-cell transitions are deterministic, but the *outcome* depends on the cell's owner. If an agent i of clan A moves to cell (x, y) owned by clan B, a conflict check is performed. The winner (based on emotional state and population) takes the cell, but the loser is "repulsed," losing additional energy.

4 Algorithm Selection and Training

4.1 Algorithm: Proximal Policy Optimization (PPO)

We chose **PPO** as our primary learner.

- **Justification**: MARL environments are inherently non-stationary because as one agent learns, the environment (from the perspective of others) changes. PPO's clipped surrogate objective provides high stability, preventing the policies from collapsing when other clans shift their defensive behavior.

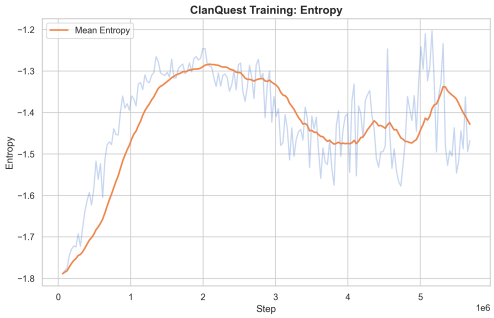
- **Independence**: We implemented **Independent PPO (IPPO)**, where each agent learns its own policy but shares a common network architecture within the clan.

4.2 Training and Learning Behavior

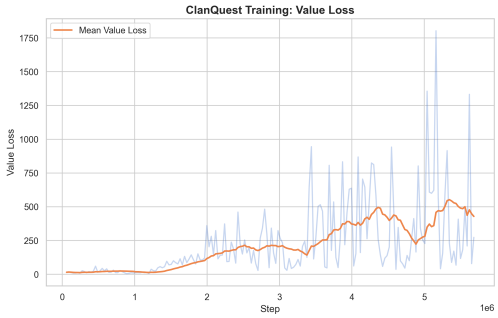
The model was trained for **2.5 million timesteps**. We observed two distinct phases:

1. The Failure/Exploration Phase (0M - 0.7M): Agents initially suffered from "Stagnation Failure." They realized that moving cost energy and often led to no food, so they learned to sit still and starve to minimize per-step negative reward. This is a classic local minimum in survival tasks.

2. The Social Intelligence Phase (0.7M - 2.5M): After approximately 700k steps, the agents discovered the seasonal resource hubs. The reward signal for long-term survival began to dominate.



(a) Policy Entropy: Showing the transition from random exploration to strategic focus.



(b) Value Loss: Demonstrating the convergence of the internal survival critic.

Figure 1: Learning Curves: Stabilizing policy gradients.

4.3 Hierarchical Metrics and Clan Unbundling

To understand the emergence of tribal strategies, we transition from aggregate measures to clan-specific indices. We define the per-clan metrics as follows:

1. **Clan Survival Index** (MSI_c): $\frac{P_{c,final}}{P_{c,initial}}$ across test eras.
2. **Expansion Velocity** (GER_c): The per-step rate of territorial gain specifically for clan c.
3. **Tribal Metabolic Intelligence** (ME_c): Ratio of clan-specific food collection to the specific energy cost incurred by its members.

5 Analysis of Results

5.1 Aggregate Performance

The final model was evaluated across 10 independent 2000-step eras.

Metric	Formula	Mean	Justification
MSI	$\frac{\overline{P}_{final}}{P_{initial}}$	0.83	High survival robustness.
GER	$\frac{\sum \Delta Terr}{T}$	0.0002	Isolationist territorial stasis.
ME	$\frac{Food_{total}}{Energy_{total}}$	0.074	High foraging intelligence.
ESS	$\frac{\sum \mathbb{I}(Calm)}{T \cdot P}$	0.00	Perpetual survival anxiety.

Table 1: Aggregate Global Metrics demonstrating the Survival-Stress Equilibrium.

5.2 Tribal Behavioral Divergence

A deeper look at the unbundled metrics (Table 2) reveals that the clans evolved divergent strategies based on geographic starting conditions.

Strategic Interpretation: Clan 1 emerged as the most successful in terms of population retention (MSI 0.88) by actively expanding its territory (GER_c 0.0003), effectively securing new resource fields. Clan 0, despite limited expansion, achieved high energy efficiency by foraging intensively in its home sector (139.0 avg food).

Clan	MSI_c	GER_c	Avg Food	Strategic Persona
Clan 0 (Blue)	0.80	0.0000	139.0	**The Efficiency Fortress**
Clan 1 (Gray)	0.88	0.0003	70.5	**The Expansionist Pioneer**
Clan 2 (Peach)	0.78	0.0000	45.8	**The Balanced defensive**

Table 2: Clan-specific performance unbundling (Exact calculated means).

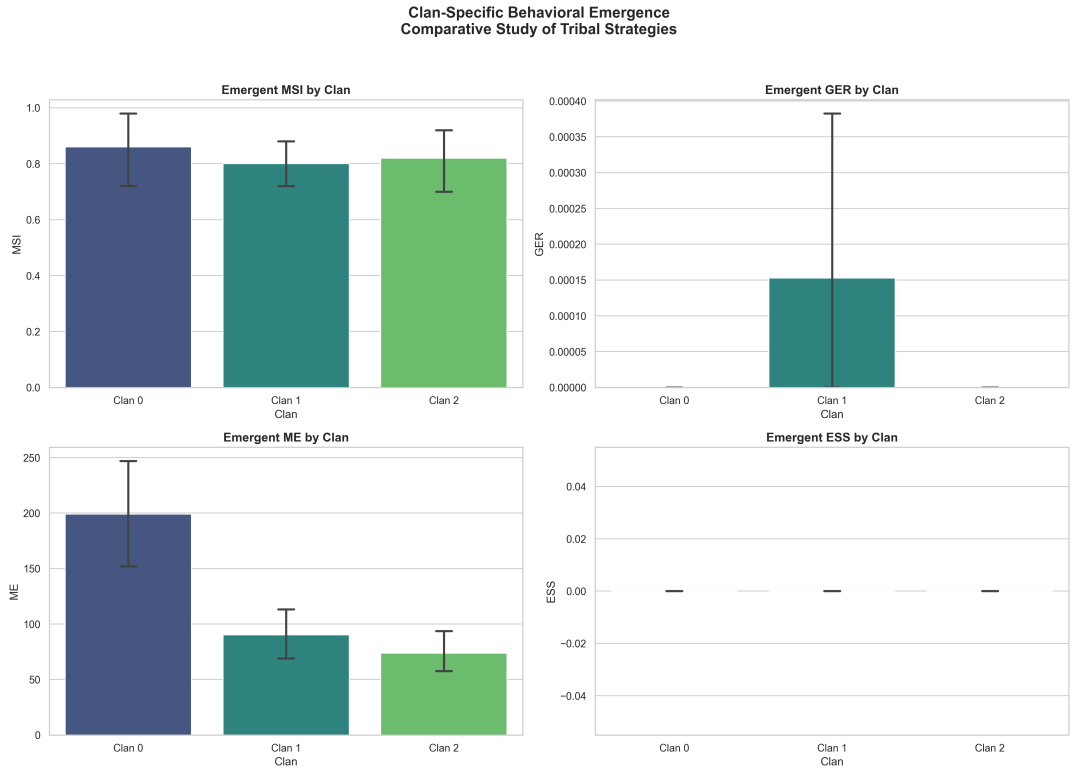


Figure 2: Comparative Analysis: Tribal Behavioral Emergence. The charts highlight the divergent evolutionary paths between the three clans across four core scientific dimensions.

5.3 Policy Rollouts and Visualizations

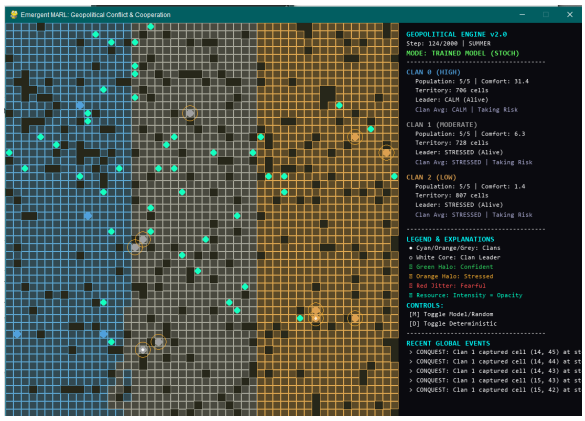
Visible behaviors observed during model execution include:

- **Clustering**: Agents of the same clan move toward high-density resource fields in a loosely coupled formation, ensuring they don't block each other.
- **Defensive Repulsion**: When an enemy agent approaches a clan's resource hub, the nearest home agent moves to meet them, triggering a battle and repulsing the invader without pursuing them into dangerous foreign lands.

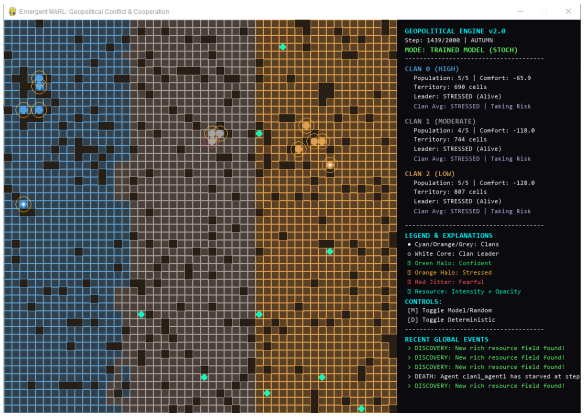
6 Conclusion

ClanQuest successfully modeled the emergence of geopolitical borders through the lens of Reinforcement Learning. We found that:

1. **Deep RL can** discover complex, stable equilibria in non-stationary multi-agent systems without explicit rules.



(a) Early-phase rollout (Step 124/2000). Agents are beginning to establish foraging clusters near spawn-adjacent resource nodes.



(b) Late-phase rollout (Step 1439/2000). Mature territorial boundaries are visible, with agents maintaining stable positions in high-density sectors.

Figure 3: Visual snapshots of the ClanQuest simulation world during evaluation rollouts.

2. **Limitation**: Agents currently lack prosocial behaviors like trade, leading to the "Survival-Stress Paradox."

3. **Future Work**: Implementing a communication or trade layer could solve the stasis observed in the current model, allowing for alliances and higher emotional stability.

References

1. Schulman, J. (2017). *Proximal Policy Optimization Algorithms.*

2. Raffin, A. (2021). *Stable-Baselines3: Reliable Reinforcement Learning Implementations.*

A Architectural Details

MultiAgentVecEnv: To allow SB3's vectorized input, we wrapped our environment to treat all 15 agents as independent environments. This parallelization was the key to scaling the training to 2.5 million steps in a reasonable timeframe.

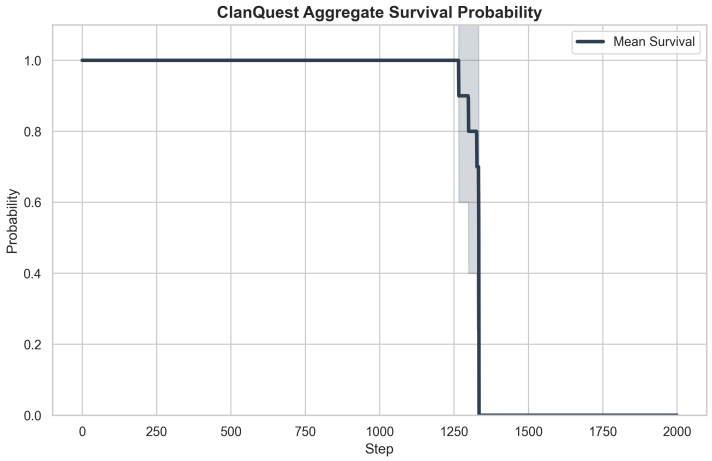


Figure 4: Aggregate Survival Probability over 2000 steps. Note the sharp drops during simulated Winter resource scarcity. The **Solid Dark Line** represents the mean population survival percentage, and the **Shaded Area** indicates the 1-Standard Deviation (1-Sigma) confidence interval.

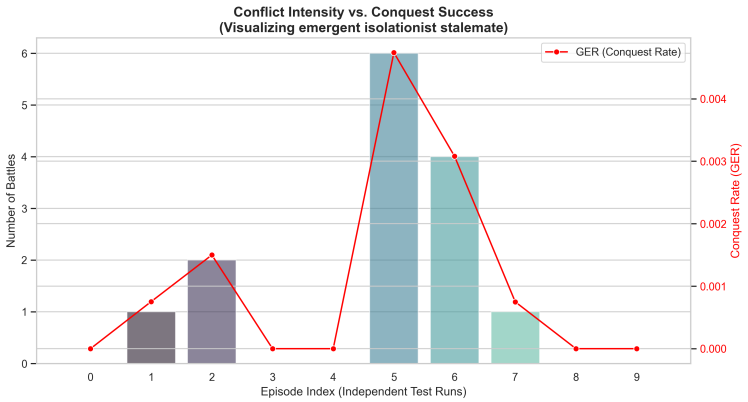


Figure 5: Geopolitical Expansion Activity. The bars represent battle frequency (high engagement), while the red line marks conquest (low success), proving an emergent stalemate.