Multi-Agent Reinforcement Learning for Wildfire Suppression: The Efficacy of Dense Local Rewards vs. Sparse Global Penalties

Abdul Rahman Hussain Siddique (ah244)
Anirudh Ramesh (aramesh8)
Yashwanth Pulimi (ypulimi)

November 22, 2025

Abstract

This study explores Multi-Agent Reinforcement Learning (MARL) for coordinated wildfire suppression using the `FirecastRL` environment. We hypothesized that a cooperative reward structure, penalizing agents for global fire spread, would incentivize superior containment strategies compared to a greedy, individualistic reward structure. Contrary to our hypothesis, our experiments demonstrated that the "Greedy" agents, motivated by dense local rewards for extinguishing fires and penalties for wasted resources, significantly outperformed the "Cooperative" team. The Greedy strategy reduced average total fire damage by approximately 39% compared to the Cooperative baseline. Our analysis suggests that in high-dimensional spatial environments with centralized control, dense cause-and-effect signals (local hits) are more effective teaching tools than sparse, noisy global signals (total fire size), mitigating the credit assignment problem.

1 Introduction

Wildfire suppression is a complex spatial-temporal problem requiring rapid decision-making and resource management. While standard Reinforcement Learning (RL) has shown promise in single-agent settings, multi-agent coordination remains a significant challenge. A core question in MARL design is the balance between individual incentives and collective goals.

We initially hypothesized that agents trained with a **Cooperative Reward** (penalizing the total number of burning cells) would learn emergent behaviors like perimeter containment, leading to better long-term outcomes. We contrasted this with a **Greedy Reward** (rewarding individual extinguishment), which we expected to yield "selfish" and myopic behavior.

2 Methodology

We utilized the `FirecastRL` environment, simulating wildfire spread on a 160x240 grid. We modified the environment to support a centralized controller managing three helicopter agents simultaneously.

2.1 Algorithm

We employed Proximal Policy Optimization (PPO) from the Stable-Baselines3 library. The architecture used a `MultiInputPolicy` to handle the dictionary observation space, which included the fire map grid and agent coordinates.

2.2 Experimental Setup

We designed two distinct reward functions to test our hypothesis:

- **Experiment A (Cooperative):**
  - *Global Penalty:* -0.1 for every cell currently burning (cumulative pressure).
  - *Team Reward:* +10.0 for every cell extinguished by the team.
  - *Waste Penalty:* -2.0 for dropping water on empty ground.

- **Experiment B (Greedy):**
  - *Global Penalty:* None (Agents are indifferent to fire size).
  - *Individual Reward:* +10.0 for successful extinguishment.
  - *Waste Penalty:* -2.0 for dropping water on empty ground.

Both models were trained for 100,000 timesteps on an NVIDIA T4 GPU.

3 Results

The results of our evaluation contradicted our initial hypothesis. The Greedy agents consistently outperformed the Cooperative team across all key metrics.

3.1 Damage Control

As shown in Figure 1, the Greedy strategy resulted in significantly lower total damage.

- **Cooperative Mean Burnt:** 1600.9 cells
- **Greedy Mean Burnt:** 973.6 cells

This represents a **39.2% reduction** in forest loss achieved by the Greedy agents.

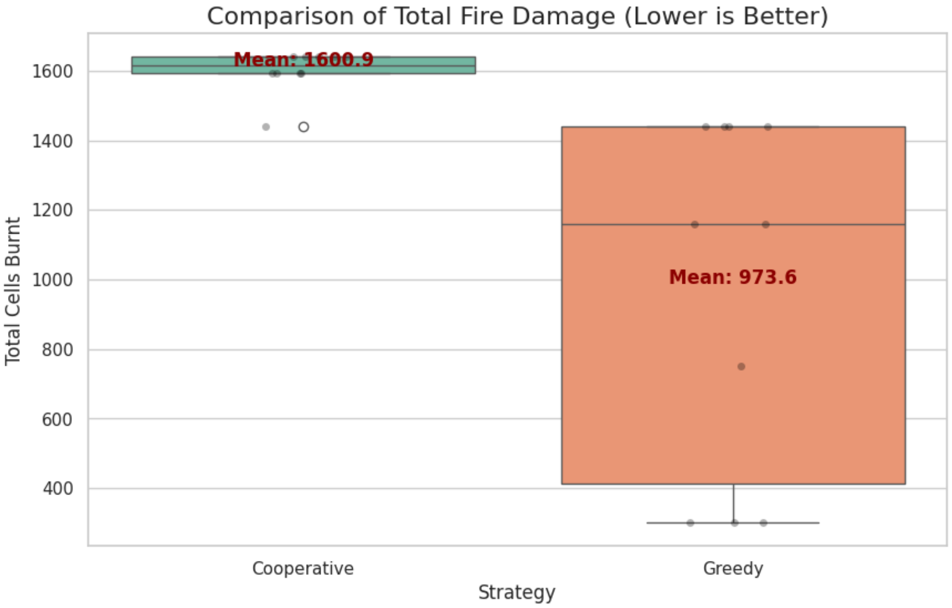

Figure 1: Distribution of Total Burnt Cells (Lower is Better). The Greedy strategy shows a significantly lower median and mean damage.

3.2 Containment Speed

Figure 2 illustrates the active fire size over time. The Cooperative agent (blue line) struggled to reduce the fire size, often allowing it to grow before stabilizing. In contrast, the Greedy agent (orange line) demonstrated a rapid and aggressive suppression strategy, driving the fire count to zero much faster.

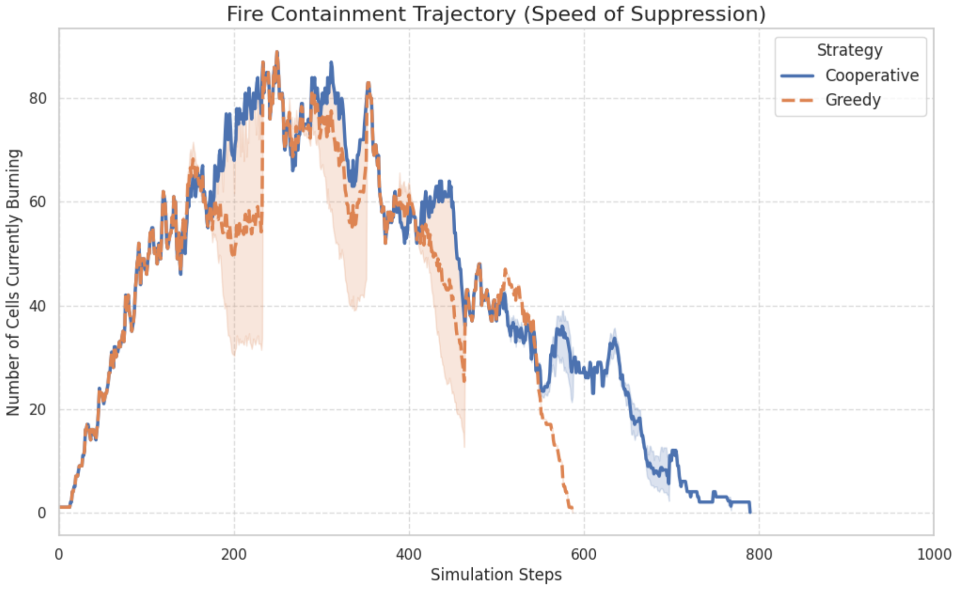

Figure 2: Fire Containment Trajectory. The Greedy agent (orange) reduces active fires to zero much faster than the Cooperative agent (blue).

3.3 Efficiency Analysis

We analyzed the relationship between episode duration and total damage (Figure 3). The Greedy agents cluster in the bottom-left quadrant, indicating they are both faster and more effective. The Cooperative agents often extended episodes without achieving better containment, suggesting inefficient loitering or hesitation.

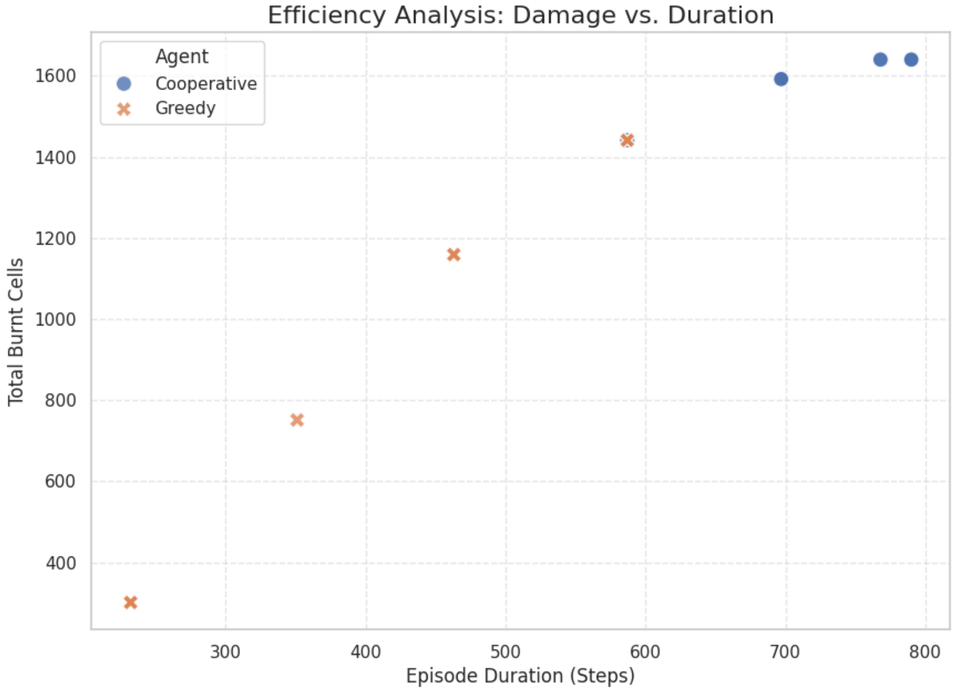

Figure 3: Efficiency Scatter Plot (Damage vs. Duration). Greedy agents cluster in the optimal bottom-left region.

3.4 Risk Profile

The Kernel Density Estimate (Figure 4) highlights the reliability of the strategies. The Greedy distribution is shifted entirely to the left, indicating that even its ”worst” runs were often better than the Cooperative agent’s ”average” runs.

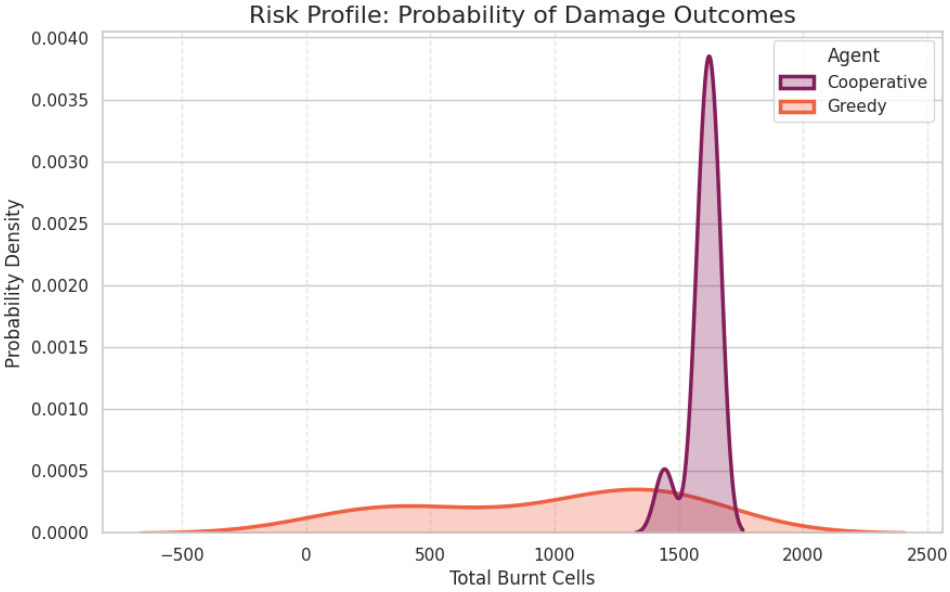

Figure 4: Risk Profile. The Greedy strategy (orange) consistently yields lower damage outcomes.

4 Discussion and Failure Analysis

Why did the Cooperative strategy fail despite having a technically ”aligned” objective (minimizing fire)?

1. **Signal-to-Noise Ratio:** The Cooperative penalty (`-0.1 * burning_cells`) fluctuates based on fire physics, independent of the agent’s immediate actions. This creates a noisy reward signal. The agent struggles to correlate a specific action (e.g., ”Move Left”) with a marginal change in a massive global penalty.

2. **Credit Assignment:** In a global reward setting, if Agent A puts out a fire but Agent B misses, the team reward is ambiguous. The Greedy reward avoids this by providing immediate, clear feedback: ”You hit -¿ +10” or ”You missed -¿ -2”.

3. **Learned Helplessness:** If the fire grows large early in the episode, the Cooperative agent accumulates massive negative rewards regardless of its actions. This can lead to a local optimum where the agent essentially ”gives up” or adopts safe but ineffective behaviors to avoid further penalties.

5 Conclusion

This project demonstrates that in centralized MARL applications for wildfire suppression, **dense, action-contingent rewards** (Greedy) are superior to **sparse, state-contingent penalties** (Cooperative). While counter-intuitive, incenting agents to aggressively ”hunt” fires individually creates a more robust aggregate behavior than punishing them for the collective state of the environment. Future work could explore ”hybrid” rewards that blend these signals to balance precision with perimeter defense.