Emergent Altruism in Multi-Agent Reinforcement Learning

Introduction

Many papers that concern themselves with AGI also consider what multiple AGIs in a singular system may come to look like. Are resources able to be shared? Do they require a common goal? In this project, I sought to investigate a scaled-down version of this by comparing two environments, one encouraging competition, the other encouraging cooperation. Specifically, the hope was to see these rudimentary AI exhibit altruism: behavior that benefits another at its own expense.

Methods

The "system" I've chosen to simulate this situation is the game Snake. Each agent is a snake, representing an AI, competing for the shared resources of fruits that pop up at a random location when the last one is eaten.

The game is how most are familiar with it, snakes are given a point every time they consume a fruit, and immediately die upon colliding with the wall, themselves, or another snake.

In the background, both the competitive and cooperative groups of agents are given the following information:

- The risk of collision in an immediately lookahead distance
- The direction of the current fruit on screen
- The direction of other snakes

Both groups of agents are also slightly rewarded for moving towards the current fruit and slightly punished for moving away from the current fruit. They are also severely punished for dying under any conditions.

However, the difference comes when looking at the reward for consuming fruit.
In the competitive group of agents, only the snake that consumed the fruit gets a reward.
In the cooperative group of agents, *all* snakes get rewarded when just one consumes a fruit. Note that this reward is unrelated to the final “score” metric with which I evaluate the success of these agents.

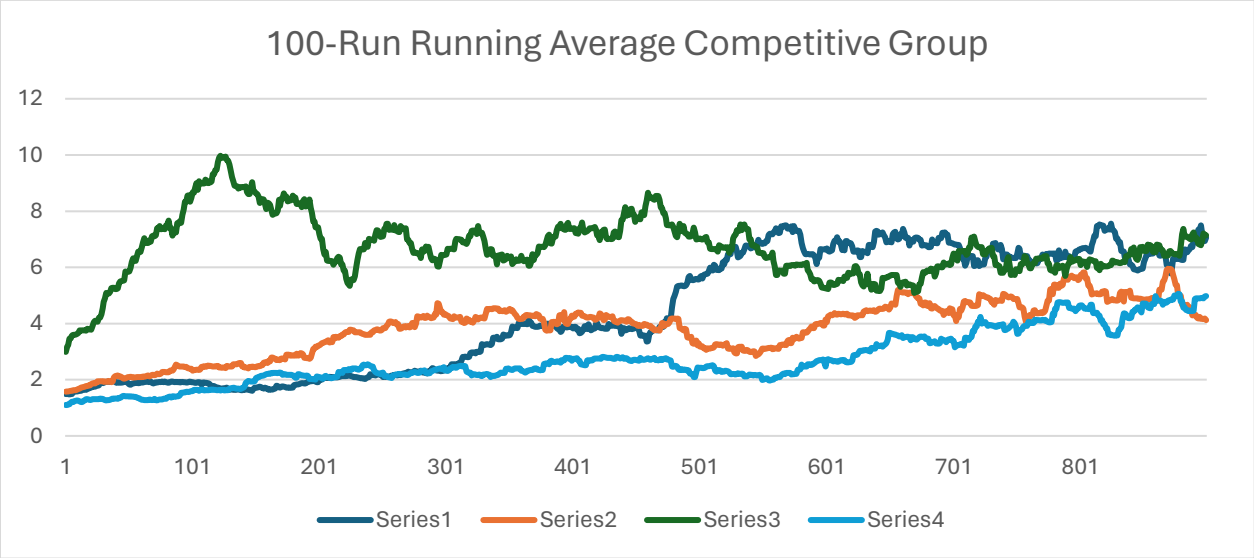
These states and reward and punishment structures are turned into actions by means of a Deep Q-Network (DQN). A DQN is a neural network architecture used in reinforcement learning that approximates the optimal action-value function, $Q^*(s,a)$ where s is the state and a is the action. In this case, s is a set of twelve bits that represent the state above, and there are four actions possible, corresponding the direction of arrow key movement.

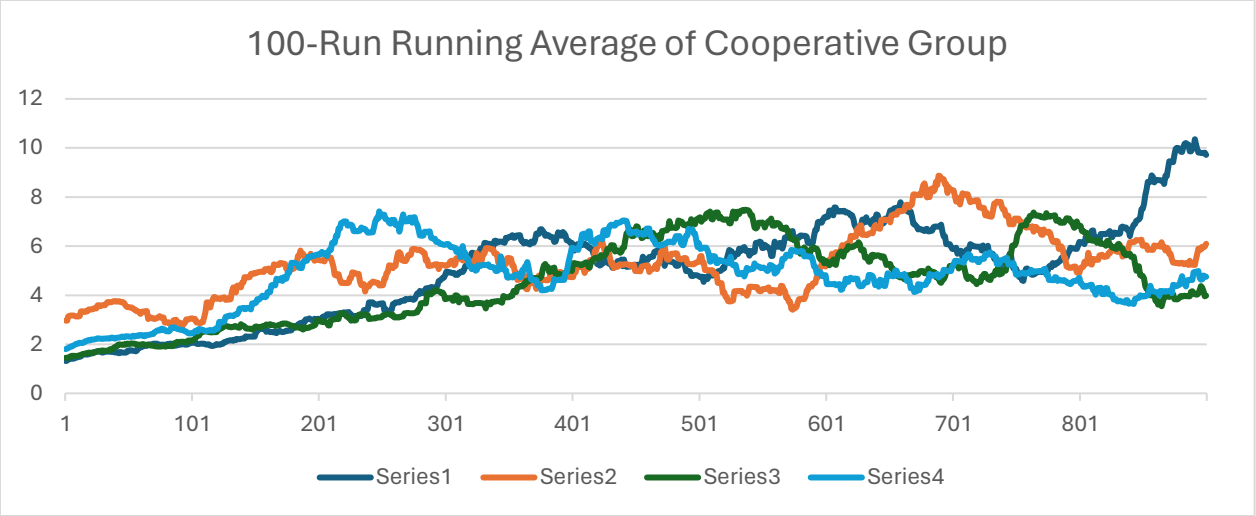
The network takes the state s as input and outputs a Q-value for each possible action. During training, the DQN uses experiences (state, action, reward, next state) to minimize the error between the predicted Q-value and the target Q-value, which is the reward plus the discounted maximum Q-value of the next state, with use of a loss function.

The metric that I will be using for actual success of this is the "score" of each of the snakes over time, ie the length of the snakes.

Results

My hypothesis for both groups is that we would see one agent come to outperform the others over time. This is inevitable, since increased survival means that that agent is trained for longer, meaning that it learns faster, makes better decisions, and survives even more. Thus, success begets success.
However, I expect to see that even more prevalent in the competitive group, with one snake rising much higher compared to the rest. In contrast, I expect for the cooperative group to have more equal performance, and for the sum of the agents' scores to be greater than that of the competitive group.





The two graphs show the 100-run running average across 1000 total runs of training for 4 agents in the competitive and cooperative groups.

The results came to be a surprise to me to say the least. While Snake 3 in the competitive group showed a strong lead for a long time, after a fair amount of time, all 4 Snakes converged to similar performance. On the other hand, the cooperative group showed more similar performance throughout the training, but nearing the end, Snake 1 showed a massive lead, but this may be chalked up to an outlier that would've converged in more time.

However I cannot say with confidence from the data or the actual video of performance if there was any altruism actually displayed in the cooperative group, or if any differences in performance were a result of the different reward structures.