1 Introduction

In certain environments, the interests of an individual agent conflict with the collective interest. This is known as a social dilemma. Leibo et al. outline three canonical two player social dilemmas in matrix form: Chicken, Stag Hunt, and Prisoner’s Dilemma. [2] In each game agents make simultaneous decisions and then rewards are distributed. Figure 1 shows the payoff matrix for each variation. Observe that mutual cooperation maximizes collective reward. However, agents also have a rational motivation to defect. In Chicken, for instance, agents can increase their reward by defecting.

<table>
<thead>
<tr>
<th>Chicken</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3, 3</td>
<td>1, 4</td>
</tr>
<tr>
<td>D</td>
<td>4, 1</td>
<td>0, 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stag Hunt</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4, 4</td>
<td>0, 3</td>
</tr>
<tr>
<td>D</td>
<td>3, 0</td>
<td>1, 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prisoners</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3, 3</td>
<td>0, 4</td>
</tr>
<tr>
<td>D</td>
<td>4, 0</td>
<td>1, 1</td>
</tr>
</tbody>
</table>

Figure 1: Payoff matrices for common forms of a two player simultaneous social dilemma.

Each game is played repeatedly. Rewards are accumulated over time. Agents must make a choice between exploiting others for immediate payoff at the expense of the long-term collective reward. A successful strategy is both temporally dependent and requires an accurate model of the other agent. Humans, as social beings, are surprisingly adept at solving these dilemmas. In general, individuals instinctively find cooperative solutions while preventing free-riders from taking advantage. Important institutions like public health and national defense depend on our ability to cooperate towards a public good. Cooperation is, in part, responsible for the success of our species.

As artificial intelligence progresses in complexity, it becomes increasingly important to have successful and well-understood pro-social behavior. How is this pro-social behavior learned? How can cooperation be generalized to different environments? How do we best balance individual reward and collective interest? How do we prevent free riders while maintaining equal rewards for all agents involved? How do humans engage with cooperative artificial intelligence? My project aims to provide further insights surrounding these critical questions.
2 Project

Getting artificial agents to behave in a human-like pro-social manner is tricky. For simple games, like the Iterated Prisoner’s Dilemma, there are defined strategies that humans use. Examples include Pavlov and generous Tit-for-Tat. [1] These strategies can easily be recreated for artificial agents. However, these strategies make heavy handed assumptions about the environment and, thus, do not generalize well to more complex environments.

More successful approaches use self-play to achieve adaptable cooperative behavior. This is typically achieved through Reinforcement Learning. Reinforcement Learning allows agents to cooperate in increasingly complex social dilemma environments. Kleiman-Weiner, for instance, use reinforcement learning to coordinate multiple agents in a Gridworld environment. This requires low-level planning over spatial action to realize strategic goals. The action space is much larger than Prisoner’s Dilemma and cooperative strategies are less obvious. [7] Building on this work, DeepMind introduces two new games: Gathering and Wolfpack. Both games can be modeled as a two player partially-observable Markov Decision Process. [2] More recently, Jaques et al. introduce two new sequential social dilemmas: Harvest and Cleanup. See Figure 2. Both require numerous agents to coordinate to preserve a shared resource. [6] These variants (numerous game states, partial observability, and multiple agents) all challenge our ability to accomplish human levels of cooperative behavior.

![Figure 2: In Cleanup, apples spawn depending on the amount of waste in the river. Agents either collect apples or clean. In an ideal situation, all agents share responsibility for cleaning the river and share in the reward equitably.](image)

To accomplish cooperative behavior, I’ve broken down my project into its five distinct parts.

1. **Environment / Learning interface**: The first step is defining a suitable interface between the environment and any learning agents. This abstraction allows me to swap between different games and intelligent agents with ease. The environment will be built using the Pycolab library built by DeepMind. It provides the basis for a highly customizable Gridworld environment. I’ll spend approximately one week building an interface between the simplest possible environment and a rudimentary actor.

2. **Multiple Agents**: I will subsequently create two agents that learn to-
gether to accomplish a trivial task. Both agents will use a Deep Q Network based on the Leibo’s approach.[2] I anticipate this will take approximately two weeks.

3. **Game Design**: After exploring a rudimentary example of multi-agent learning, the next step is designing an intelligent game. There are many variations of social dilemmas. Each has unique advantages and disadvantages. The environment has a large impact on how easily cooperation is learned. I’ll rely heavily on current literature to make intelligent design choices. I believe Cleanup (Figure 2) will be a suitable choice. However, more work is required to confirm this intuition. The code for games is unavailable publicly and must be reimplemented.

4. **Cooperation**: Contingent on success of the first three steps, the bulk of the project will be spent exploring and refining the cooperative mechanism between two agents in the constructed social dilemma environment.

5. **Communication (Stretch)**: With any remaining time, I plan to add a communication channel between agents to aid in cooperation. Jaques et al. use a reinforcement technique that trains agents to provide relevant information over such a channel. They also endow each agent with a desire (reward) to causally influence the other agent. As a result of these modifications, agents learn to communicate useful information about the environment that other agents can act upon.[6] I aim to replicate their results. Hopefully pro-social agents with communication can better cooperate than pro-social agents without this ability. This goal is entirely contingent on first achieving successful cooperation. Should cooperation fail, I’ll instead invest my time diagnosing why.

3 Deliverables

At the end of this project I will provide all relevant code, the weight values for my trained networks, and a final report documenting my results.

4 Cooperation

Ngan Vu, Zen Tang, and Simon Mendelsohn will be working on pro-social interactions in conjunction with my project. They will be using the same interface to enable quick comparison. However, my results and code will be clearly separated from any work and coding that they do in parallel.
References


