Fictitious Co-Play for human-agent collaboration [1][2][3]

TU Delft - CSE3000
Nathan Ordonez
n.a.ordonezcardenas@student.tudelft.nl
Supervisors: Frans Oliehoek, Robert Loftin

[1] Background

Purpose

Autonomous agents will inevitably work alongside humans one day. Their physical strength and processing power will enable new possibilities for human endeavours.

Through **collaborative** games, researchers aim to find the **best reinforcement learning algorithms** to control those agents.

Related work

Some of those **algorithms** include (see top-left figure):

- *Self-Play* **(SP)**, where an agent trains with a copy of itself until it reaches the desired level of skill.
- *Population-play* **(PBT)**, where a population of agents with varying hyperparameters are trained with each-other, and evaluated. Then, a subset of best-performing agents is picked and the process is repeated.
- *Behavioral cloning play* **(BC)**, where a large amount of data is recorded from two humans playing with each-other, and an agent is trained to copy the recorded data. The goal is for it to *play as a human would.*

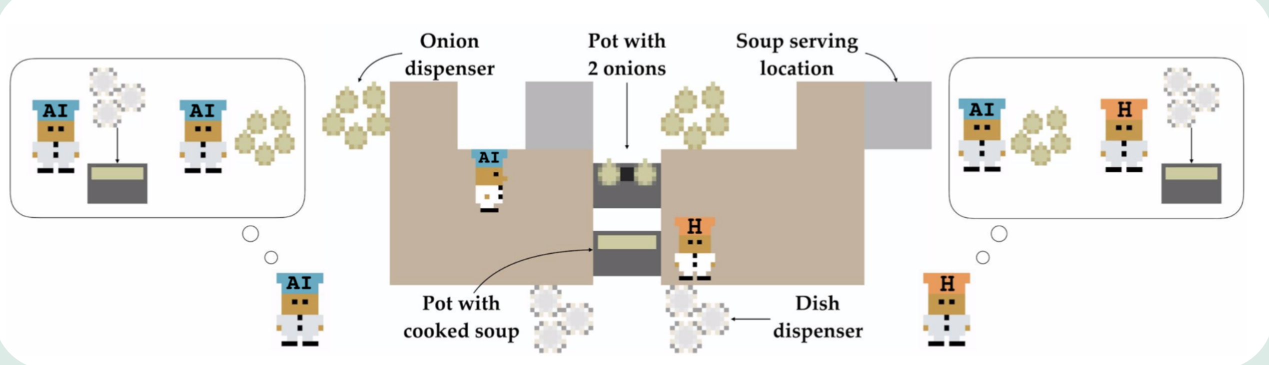
The **problem** with those methods is that they either do not learn to generalise well to human beings, or they human-human data that is expensive to acquire at scale.

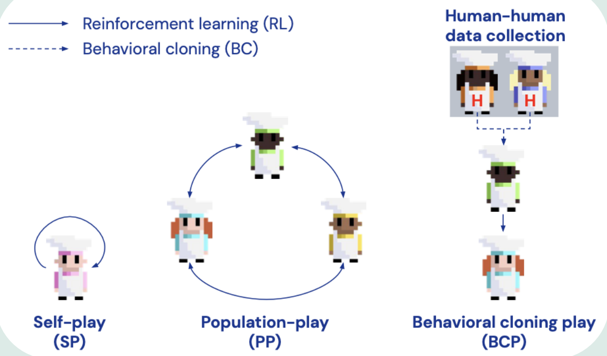
Problem statement

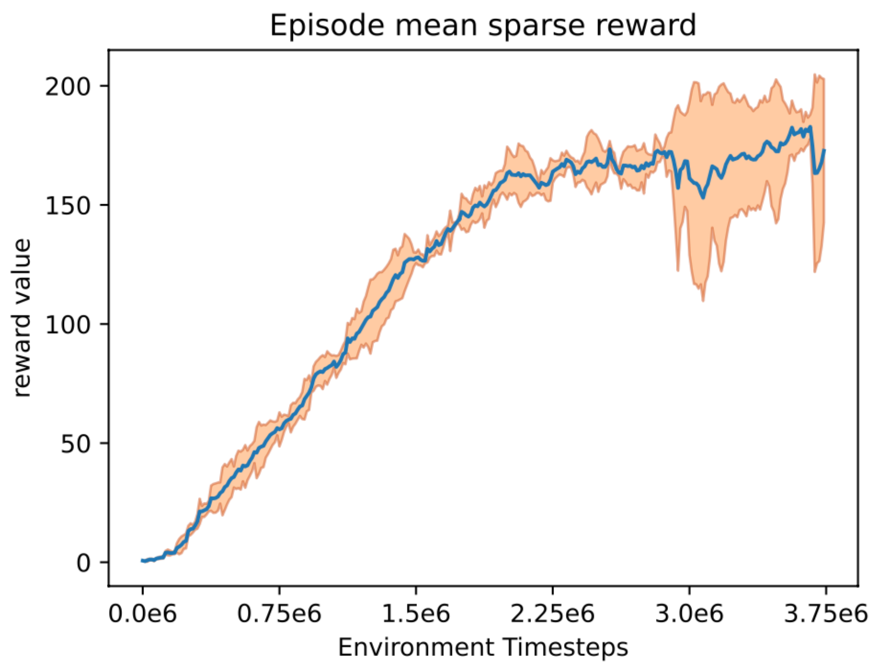
In this work, we tackle the challenge of the **Overcooked** environment [1], based on the "Overcooked!" environment of the same name. Agents must **collaborate** in order to deliver as many onion soups as possible.

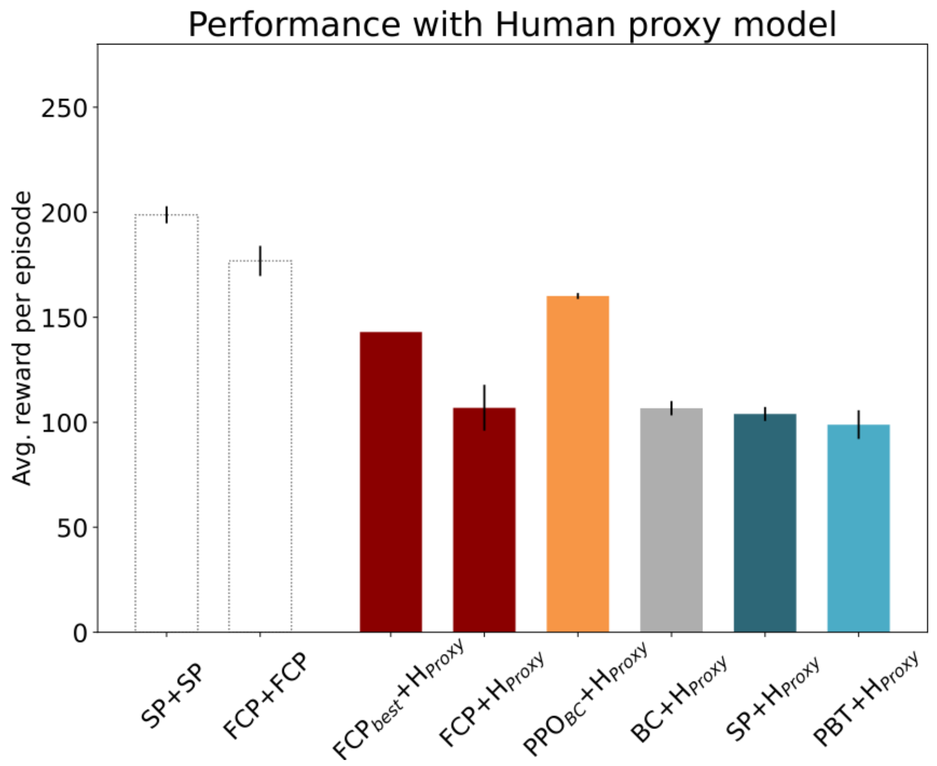
In particular, we focus on a state-of-the-art technique called "Fictitious Co-Play" [3] which was ideated by a paper by DeepMind. **In our work, we contribute to research by attempting to reproduce their results at a much smaller scale using accessible amounts of computation** to see whether the method shows great advantages over two methods (SP, PBT, BC) like its original paper found.

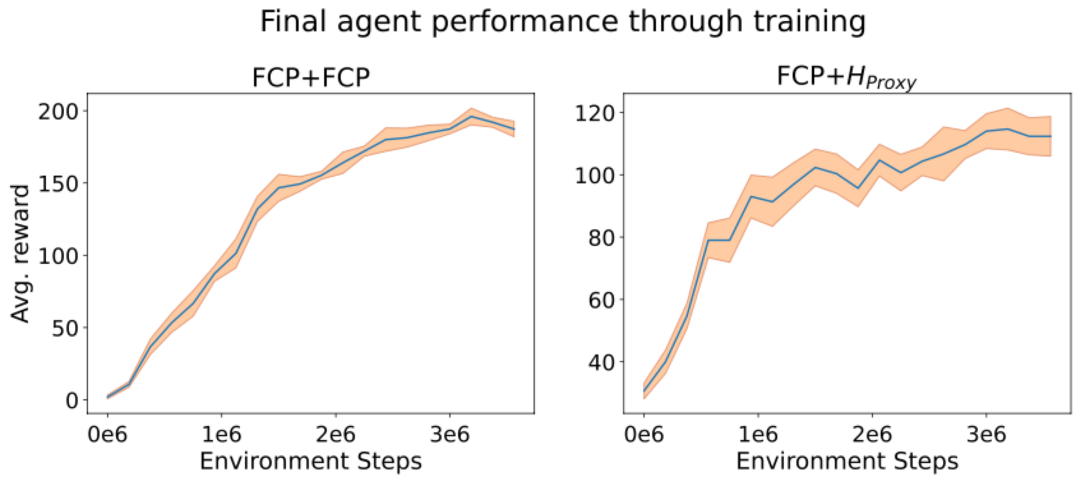
So the question is: "**Can the FCP method generate agents that adapt well to human beings?**"




Reinforcement learning (RL)
Behavioral cloning (BC)
Human-human data collection
H
Self-play (SP)
Population-play (PP)
Behavioral cloning play (BCP)



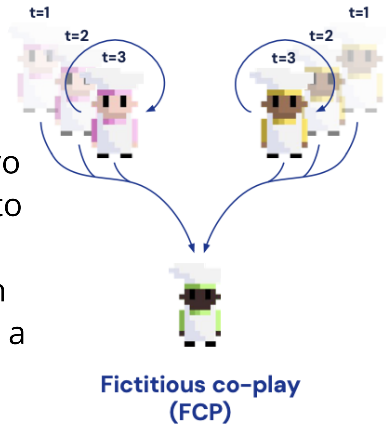




[2] Method

The FCP method explained in two steps:

1. Multiple **Self-Play** agents are trained, and two checkpoints are saved during their training (to have agents with varying levels of skill).
2. A **final agent** is initialized, who will train with each Self-Play agent and their checkpoints in a number of episodes.



How this method works:
The basic idea of FCP is that in order to learn to play with humans without actually training with them, the agent must train with a **diverse** set of training partners.
The trick is to use the fact that Self-Play agents tend to optimize towards **one playing style**, in order to build a population of agents **diverse** in style, yet high-performing with the right teammate. A great learning environment for the final agent.

[3] Results

Experiments

All the following FCP data are averaged over three seeds.

Agent training (top-left curve)
First, we show our final agent's reward training curve, trained with 8 Self-Play population agents and checkpoints during 60 training episodes, and a 75% confidence interval. We see that the agent learns successfully, and its training reward becomes unstable after 3 million steps.
Comparison (top-right bars)
Now that we know our agent is training properly, we evaluate it with a so-called "human proxy" (H_{proxy}) which is a behavioral cloning agent trained on public human data. We show the evaluation score averaged over 40 evaluation rounds. The method on average slightly outperforms the other methods (SP, PBT, BC), and its best score during training is shown to be significantly high. Note that the PPO_{bc} agent has a higher score because it had access to an agent similar to the human proxy, trained on a similar dataset.
Evaluation through training (bottom curves)
In order to see how the final agent improves during its training, we show the evaluation scores for twenty checkpoints of the agent (avg. of three seeds), saved during its training. We notice that the agent continues to train well, and that the instability shown in the "**Agent training**" plot does not translate to a loss in performance. **This supports the idea that the agent continues to learn from its diverse population beyond its original high-growth phase**.

Conclusions

Our results do not correspond with the doubling of performance found in the orignal FCP paper, however they do correspond with another paper that re-evaluated the FCP method, just like we did [2]. Overall, we have shown that the agent indeed shows an improvement over other methods, thus it can train agents that generalise better to playing with human beings. We also suggest that there is much room for improving the method further.

Citations:
[1] M. Carroll, R. Shah, M. K. Ho, T. L. Griffiths, S. A. Seshia, P. Abbeel, and A. Dragan, "On the Utility of Learning about Humans for Human-AI Coordination," Jan. 2020. Number: arXiv:1910.05789 arXiv:1910.05789 [cs, stat].
[2] R. Zhao, J. Song, Y. Yuan, H. Haifeng, Y. Gao, Y. Wu, Z. Sun, and Y. Wei, "Maximum Entropy Population-Based Training for Zero-Shot Human-AI Coordination," May 2022. Number: arXiv:2112.11701 arXiv:2112.11701 [cs].
[3] D. J. Strouse, K. R. McKee, M. Botvinick, E. Hughes, and R. Everett, "Collabo- rating with Humans without Human Data," Jan. 2022. Number: arXiv:2110.08176 arXiv:2110.08176 [cs].