Leveraging Cognitive Bias for Zero-Shot Human-AI Coordination

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in multiagent reinforcement learning have enabled artificial agents to coordinate effectively in complex domains; however, these agents can struggle to coordinate with humans, in part due to their implicit but inaccurate assumptions of optimal decision-making and behavior homogeneity when interacting with humans. Although we can train models to learn the best responses to human behavior given a large corpus of human-human interaction, the cost of collecting this data can be prohibitive. We show how even without such data, we can use our knowledge of biases and limitations in humans to construct a technique that can coordinate with humans. To do this, we present an approach that learns models partnered with reinforcement learning agents that incorporate human behavioral biases. We evaluate this method in the fully-cooperative game Overcooked. Our results show an improvement when incorporating this bias with methods that do not include this bias in their agent population.

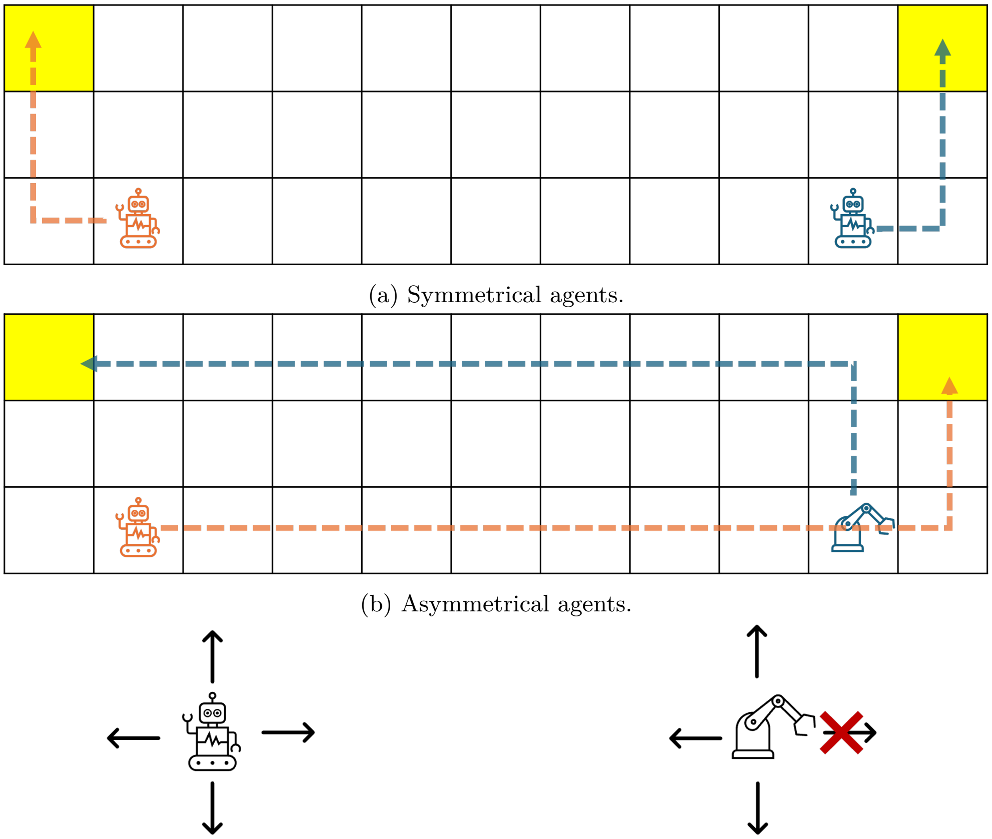
1 Introduction

We study the problem of zero-shot human-AI coordination where an agent is paired with a human in a cooperative task without prior access to data on human behavior in the task. This is a special case of the problem of zero-shot coordination (ZSC) where an agent performs a cooperative task with one or more agents without previously interacting with the agents.

Prior works in ZSC propose using reinforcement learning (RL) to train a best-response (BR) agent to coordinate with a diverse set of other agents, usually also trained with RL. This prevents the BR agent from learning a single convention to solve the problem since it has to be able to coordinate with a multitude of other agents. The challenge is to learn agent behavior that is, compatible with, or, adaptable to, any goal-driven agent. If the interacting agent is chosen at random from the set of all possible agents, all feasible actions are equally likely and adaptation is infeasible. One way to solve this issue is by limiting the deviation from optimal behavior when training the agents (Zhao et al., 2023; Lupu et al., 2021). However, this regularization of the agents is uninformed, *e.g.* Zhao et al. (2023) include a maximum entropy objective.

We know all the interacting agents are humans in zero-shot human-AI coordination. Although individual humans have different behaviors, the class of humans has some systematic biases. We explore using known cognitive and behavioral biases to generate a set of agents to help enable the BR agent to coordinate with humans (Norton et al. (2012), Ainslie & Haslam (1992)). We show that our method can achieve similar or better coordination with humans and human-like agents than methods that generate agents with unbiased diversity measures. We believe that incorporating this knowledge in our approach can help us learn agents compatible with human behavior with fewer samples.

To understand why this might happen, let us take an example of a grid-world coordination scenario in 1. In (a), there are two symmetrical agents each trying to reach one of the goal locations in yellow. The trajectory depicts one of the optimal solutions. In (b), we assume that the agents are



(a) Symmetrical agents.

(b) Asymmetrical agents.

(c) Agent action spaces.

Figure 1: Example scenario where the goal of the agents is to occupy one of the two yellow squares. In (a) the agents are symmetrical and have a straightforward optimal path to the goal closest to them. In (b), the agents have to take a circuitous path to the goal further from them due to the inability one of the agents to move right as shown in (c).

asymmetric with one agent unable to move towards the right. This leads to a different solution where the agents have to reach for the goal locations that are further away from them. Such a behavior is very unlikely to emerge if we sample diverse behaviors of symmetric agents.

Our observation is that humans and artificial agents agent are not symmetric and leveraging the behavioral biases and skill differences in a principled manner can improve human-AI coordination. Our main contributions are as follows:

1. We convert human biases into training objectives for RL agents.
2. We show improved task performance using these metrics in Overcooked as compared to other methods using unbiased diversity and other state-of-the-art methods in this domain.

2 Background

We model interaction as a multi-agent extension of the Markov decision process, know as a Markov Game $\mathcal{M}$, defined as a tuple $\langle \mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \boldsymbol{r}, \mathcal{T}, \gamma \rangle$. Here, $\mathcal{N} = \{1, \ldots, n\}$ is the set of agents, $\mathcal{S}$ is the set of joint states, $\boldsymbol{\mathcal{A}} = \{\mathcal{A}^1, \ldots, \mathcal{A}^n\}$ is the set of joint actions, $\boldsymbol{r} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times N \rightarrow \mathbb{R}$ is the reward function, $\mathcal{T} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \rightarrow [0, 1]$ is the transition function, and γ is the discount factor. At each timestep t, agent $i \in \mathcal{N}$ receives the state $s \in \mathcal{S}$, and samples an action, $a \in \pi_i(s)$, according to a policy

$\pi_i : \mathcal{S} \mapsto \mathcal{A}_i$. The expected return for a joint policy $\pi \in \Pi$ is defined as $J(\pi) = \mathbb{E}_{\pi}^{\mathcal{M}}[\sum_{t=0}^{T} r(s_t, a_t)]$. Here, we consider the fully cooperative setting, where $r_i = r_j \forall i \in \mathcal{N}, \forall j \in \mathcal{N}$.

Zero Shot Coordination (ZSC). In two-agent ZSC, the goal is to learn a policy that can enable it to coordinate with an unknown agent., $\arg\max_{\pi} \mathbb{E}_{\pi' \in \Pi} J(\pi, \pi')$. Although, Π is the set of all policies, prior work only considers goal-driven agents. In ZSC for human-AI coordination, the goal is to find a policy that can coordinate optimally with human policies, $\pi' \in \Pi^H \subset \Pi$. We are concerned with this goal in this work.

Self-Play Coordination (SP). The objective of SP is to learn a policy that can enable coordination between agents when both are following the same policy, $\pi' = \pi$. This is an easier problem to solve, however, SP policies can learn arbitrary conventions that lead to suboptimal performance when interacting with agents following a different policy, *e.g.*with humans.

Best-Response (BR). An agent is BR to another if it follows a policy that coordinates well with the known agent, $\pi_{BR} = \arg\max_{\pi} J(\pi, \pi')$, where policy π' is known. BR to a set of known policies Π' follows a policy that can coordinate with policies within that set, $\pi_{BR} = \arg\max_{\pi} \mathbb{E}_{\pi' \in \Pi'} J(\pi, \pi')$.

3 Related Work

Recently, there has been inspiring research in enabling zero-shot coordination between agents, human and AI. While, Carroll et al. (2019) provide evidence for the importance of collecting and utilizing human interaction data for learning policies that can interact with humans, others show evidence of training coordinating agents by pairing them with a large population of diverse agents (Strouse et al., 2021; Zhao et al., 2023). Our work is similar to the latter in utilizing a diverse population of agents to enable learning a policy that is adaptable to humans. However, unlike prior work, we rely on information of human behavioral biases to guide learning the agent population.

Cognitive Bias: Humans are subject to various cognitive biases that influence their decision-making processes (Simon (1997)). For example, the IKEA effect is one where individuals place a higher value on products they partially create, reflecting an increased attachment to their own efforts (Norton et al. (2012)). Humans also have inherent limitations on how quickly they can react to situational changes, and this reaction speed varies from person to person. Another cognitive bias is the tendency to prefer immediate rewards over future rewards (Ainslie & Haslam (1992)). Studying these cognitive biases and incorporating them into RL algorithm can help train bias-aware agents that are better at human-AI coordination.

4 Method

Our goal is to find a policy that can coordinate with human policies Π^H. Our method uses reinforcement learning (RL) in two stages. First, we generate a set of policies, by sampling known cognitive biases. Second, we learn an agent as a best response to the set of policies with the goal of achieving ZSC for human-AI coordination.

Prior literature finds that BR to a population of agents with some form of diversity enables coordination with other agents, including humans. Our method takes inspiration from this as well as the literature investigating cognitive and skill priors in human behavior, to train a population of RL policies. We consider the following two biases:

1. Speed Asymmetry. Humans and agents do not have the same speed of action and how quickly different humans act can vary. We model this by taking an action from a trained SP model with probability p and taking no action otherwise. This also has the added benefit of training efficiency since a population of agents with differing behavior can be produced, by varying p, while training only a single RL model.

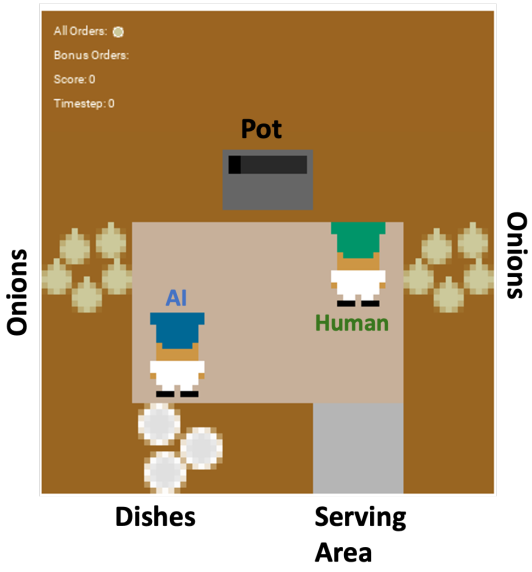


Figure 2: A human and an AI agent interact in a shared environment in the fully-cooperative game of Overcooked.

2. Time discounting. Humans often value immediate rewards over future rewards. We model this by varying the discount factor γ when training the SP model.

We use this population of policies to train the BR agent, by pairing it with randomly sampled policies from the population during the training process. Our algorithm for doing this is similar to the one used by Strouse et al. (2021)

5 Experiments

Overcooked. We utilize the Overcooked environment introduced by Carroll et al. (2019) due to its combination of strategy and motion coordination challenges, making it well-suited for deep reinforcement learning algorithms. In this setting, two agents collaborate to cook and serve soup, aiming to deliver as many soups as possible. While the original study outlines five MDPs, our preliminary experiments focus on the one illustrated in Figure 2. In this MDP, the agents' objective is to deliver soup, which involves placing three onions in a pot, cooking them for 20 timesteps, transferring the soup into a dish, and serving it. The primary challenge lies in the agents' ability to navigate the environment, interact with objects, and coordinate their strategies. The action space consists of six possible actions: `up`, `down`, `right`, `left`, `noop`, and `interact`. For training RL agents, we used the JaxMARL Overcooked environment by Rutherford et al. (2024).

We present experimental results in Table 1. Here we compare the average return per episode for three types of agents over an episode length of 400 timesteps. The Self-Play agent is a single SP agent trained for this game. The FCP' model is based on Strouse et al. (2021) but we limit the population to only two SP agents. Our method BR (Action Speed), is a best-response to SP and SP ($p = 0$). The results show us that the SP model has the lowest return, while FCP' improves its performance, but our approach performs the best while requiring less computation than FCP'.

6 Conclusion and Future Work

Although this is a work-in-progress requiring a more extensive set of experiments to validate our conclusions, the preliminary results demonstrate the potential of this idea through the increased

Method	Self-Play	FCP'	**BR (Action Speed)**
Avg. Episodic Returns	92	108	152

Table 1: Performance with proxy human. The average accumulated reward when the agents are paired with the proxy human model. Our method (in bold) was trained with two self-play models with different speed of action $p = 1, 0$.

efficiency even when including a simplistic behavioral bias. The next steps of this work will focus on:

- Implementing the proposed cognitive biases.
- Performing ablation studies to analyze how cognitive biases affect the model's performance.
- Conducting experiments involving human participants and evaluating model's performance.

References

George Ainslie and Nick Haslam. Hyperbolic discounting. 1992.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems*, pp. 5175–5186, 2019.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*, pp. 7204–7213. PMLR, 2021.

Michael I Norton, Daniel Mochon, and Dan Ariely. The ikea effect: When labor leads to love. *Journal of consumer psychology*, 22(3):453–460, 2012.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Garðar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, et al. Jaxmarl: Multi-agent rl environments and algorithms in jax. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2444–2446, 2024.

Herbert Alexander Simon. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press, 1997.

DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34: 14502–14515, 2021.

Rui Zhao, Jinming Song, Yufeng Yuan, Haifeng Hu, Yang Gao, Yi Wu, Zhongqian Sun, and Wei Yang. Maximum entropy population-based training for zero-shot human-ai coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6145–6153, 2023.