Cognitive Bias for Human-AI *ad hoc* Teamwork

Shray Bansal[*]
Georgia Institute of Technology

Jin Xu
Georgia Institute of Technology

Miguel Morales
Georgia Institute of Technology

Jonathan Streater
Florida State University

Ayanna Howard
The Ohio State University

Charles L. Isbell
University of Wisconsin-Madison

Abstract

Advancements 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 about optimal decision-making and behavioral homogeneity while interacting with humans. Although we can train models to learn the best responses to human behavior using a large corpus of human-human interaction, the cost of collecting this data can be prohibitive. We demonstrate how, even without such data, we can leverage our knowledge of biases and limitations in human behavior to develop a technique for effective human-agent coordination. To do this, we present an approach that trains an RL agent by best responding to a pool of other agents that incorporate human behavioral biases. We evaluate this method in the fully-cooperative game Overcooked. Our results show an improvement when incorporating these biases compared to methods that do not account for these biases within their agent population.

1 Introduction

We study the problem of human-AI *ad hoc* teamwork (AHT) where an agent is paired with a human in a cooperative task without prior access to data on human behavior in the task. We show that leveraging prior knowledge of human behavior in the form of skill asymmetry and cognitive bias can help us learn reinforcement learning agents that can coordinate with human-like agents while reducing training time in the fully-cooperative game Overcooked.

Prior works in AHT 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 agent. If the interacting agent is chosen at random from the set of all possible agents, all feasible actions become equally likely and adaptation is infeasible. One way to avoid this issue is by assuming that the interacting agents have the same goals but may deviate from optimal behavior [24, 11]. [19] use this assumption to train agents that can coordinate with other agents optimizing for the same rewards, while including partially trained agents to introduce skill diversity, and others rely on statistical metrics like maximum entropy [24] in the objective to induce diversity in goal-driven behavior while reducing the number of agents sampled.

[*]Corresponding author sbansal34@gatech.edu

NeurIPS 2024 Workshop on Behavioral Machine Learning.

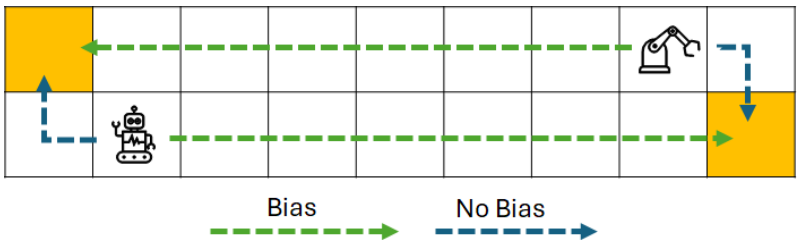

Figure 1: Example cooperative game where the goal is to occupy both yellow squares. Without bias, the optimal trajectory is to move to the nearest yellow square, but if either agent is biased against moving vertically, the green trajectory becomes optimal.

In our problem, human-AI AHT, we know that the agents will be interacting with people and so we want to leverage known properties of human behavior. Although individual humans have different behaviors, the class of humans has some systematic biases. Inspired by research in cognitive and social sciences [13, 1], we use cognitive biases to generate a set of agents to help enable the BR agent to coordinate with humans. We show that our method can achieve similar or better coordination with humans and human-like agents than methods with other diversity metrics using fewer samples.

Let us take a grid-world coordination scenario in Fig. 1 where two agents have the joint goal of reaching the yellow squares. If the agents are symmetrical and can move in all four directions, the blue trajectory represents an optimal solution. However, if one of the agents has a strong bias against moving vertically, the green trajectory becomes optimal. Due to its suboptimality in the symmetric case, it is unlikely to emerge by sampling diverse behaviors. Through this simple example we hope to illustrate how even simple biases can have large coordination effects.

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

1. Present an approach to train RL agents capable of coordinating with humans by incorporating human cognitive biases into a group of RL agents.
2. Show improved task performance and training efficiency in Overcooked as compared to other methods that do not utilize these biases.

2 Related Work

There has been growing recent research in *ad hoc* teamwork [12] and the related problem of zero-shot coordination (ZSC) [8]. We review some prior work from the lens of game theory.

Game Theory. In pure coordination games with multiple equilibria, it is in the agents' best interests to coordinate on a single equilibrium, but this coordination is challenging without prior agreement leading to the equilibrium selection problem [7]. This is one of the primary challenges in AHT, since agents are fully cooperative but lack prior interaction or agreement. Solutions to equilibrium selection can be categorized into two types: (1) solutions relying only on endogenous game information, and (2) solutions that also incorporate exogenous information about agents.

AHT methods using only endogenous information include: handling game symmetries in Hanabi by learning permutation-robust policies [8], learning best-response agents capable of handling multiple equilibria by pairing them with a population of independently trained SP policies [19], and, increasing policy diversity in the SP population, *e.g.*, by including a maximum entropy objective [24].

Humans seem to rely on exogenous information, *e.g.*, [15] showed that humans were able to coordinate significantly better than chance when playing a coordination game where agents aim to choose the same side of a coin without prior interaction. Examples of work in human-AI coordination include: using information about human behavior and social norms to group Nash equilibria and adapt online to human behavior in a table-top manipulation task [3], and, leveraging information about human bias to generate multiple event-based reward functions and learn a BR policy in Overcooked [23]. Our approach also utilizes exogenous information in the form of systematic biases in human behavior.

However, unlike [23], which sample reward functions based on pre-defined game *events*, we sample policies based on human behavioral traits without introducing a new reward structure.

Cognitive Bias: After scientists began formalizing human behavior as rational actors ([21]), they also began describing systematic human deviations from these classic notions of rationality [20, 9]. These systematic deviations, termed cognitive biases, have been identified in myriad environments and contexts, and have been used to introduce frameworks like bounded rationality [17, 5], ecological rationality [2], and resource-rational analysis [10]. Understanding human cognition as optimal and general under limitations of time, computation and communication ([6]), might help us formalize and introduce these patterns of human behavior as inductive biases in AI systems that need to coordinate with humans. Towards this goal, we take into account two human limitations: (1) limitation on human reaction speed to situational changes, and, (2) preference for immediate over future rewards [1].

Availability of Data. Human behavior data can help train agents that successfully coordinate with humans, as shown by [4]. However, collecting this task-specific data for every scenario that AI agents will interact with humans is impractical, and human behavior can evolve over time, or, vary when interacting with AI versus other humans. We aim to identify task-invariant properties of human behavior applicable across domains. Even when human data is available, our method can be used as a prior for agent policies, potentially leading to more robust AI agents as observed by [22].

3 Background

We model interaction as a two-agent common-payoff Markov game, $\mathcal{M}$, defined as a tuple $\langle \mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \boldsymbol{r}, \mathcal{T}, \gamma \rangle$. Here, $\mathcal{N} = \{1, 2\}$ is the set of agents, $\mathcal{S}$ is the set of joint states, $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_1, \mathcal{A}_2\}$ is the set of actions for each agent, $\boldsymbol{r} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ is the common reward function,, $\mathcal{T} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0, 1]$ is the transition function, and $\gamma \in [0, 1]$ is the discount factor. At each timestep t, agent i receives the state s_t, and samples an action, $a_{i,t} \sim \pi_i(s)$, according to a policy $\pi_i : \mathcal{S} \mapsto \mathcal{A}_i$. We define the expected return for a joint policy as $J(\pi_1, \pi_2) = \mathbb{E}^{\mathcal{M}}_{a_t \sim (\pi_1, \pi_2)}[\sum_{t=0}^{T} r(s_t, a_t)]$, where a_t is a joint action, and, an episode goes from time 0 to T.

Ad-hoc teamwork. The goal of ad-hoc teamwork is "to create an autonomous agent that is able to efficiently and robustly collaborate with previously unknown teammates on tasks to which they are all capable of contributing as team members" [18]. We can write this as, $\arg\max_{\pi} \mathbb{E}_{\pi' \sim \Pi^{cap}} J(\pi, \pi')$, where Π^{cap} is the set of policies of capable team members.

Self-Play (SP). In self-play, the objective is to maximize the expected return by finding the optimal joint policy, $\vec{\pi}^* \in \arg\max_{(\pi_1, \pi_2)} J(\pi_1, \pi_2)$. This solution is a Pareto-optimal equilibrium because if either agent can improve the return by selecting a different policy then it will contradict the $\arg\max$. However, it fails as a solution for ad-hoc teamwork because it assumes both agents follow the same equilibrium, or compatible policies, which is not guaranteed even if both agents were trained by SP.

Best-Response (BR). In best-response, the objective is to maximize the expected return in response to a fixed policy of the other agent. We consider a policy π^B, to be BR to π, if $J(\pi^B, \pi) \geq J(\pi', \pi) \forall\, \pi' \in \Pi$. We define the BR function, $\mathcal{B}$, such that $\pi^B \in \mathcal{B}(\pi)$. Similarly, we define BR over a policy set, $\Pi_K = \{\pi^1, \ldots, \pi^K\}$ as,

$$\mathcal{B}(\Pi_K) \in \arg\max_{\pi_{BR}} \mathbb{E}_{\pi' \sim U(\Pi^K)} \left[J(\pi_{BR}, \pi') \right], \tag{1}$$

where U is the uniform distribution.

4 Approach

Our goal is to develop a method that helps an agent find policies that effectively coordinate with human behavior. Human behavior may not align perfectly with optimizing the rational self-play objectives for several reasons, such as the skill difference between humans and autonomous agents (*e.g.* bounded rationality, reaction speed), and, cognitive bias (*e.g.* hyperbolic time discounting, preference for specific sub-tasks).

We want our agent to collaborate effectively with humans, so we train our agent to respond optimally to the behaviors that humans are likely to adopt, $\pi^{\text{BR(H)}} \in \mathcal{B}(\Pi^{H*})$. Here, Π^{H*} is the unknown set

of all human policies. To derive this method we make the assumption that human behavior can be described by a set of policies, and each policy is an equilibrium for some Markov game. Our goal, then, is to learn an agent that can adapt to this human behavior, instead of trying to influence it.

For this, we use reinforcement learning (RL) in two stages. First, we find human-like self-play policies by placing constraints on the policy-space based on a subset of known human skill factors and modifying the Markov game $\mathcal{M}$ to account for cognitive biases, $\Pi^{H_{\text{bias}}}$, see Algorithm 1. Second, we train a policy as best-response to the human self-play policy set, $\mathcal{B}(\Pi^{H_{\text{bias}}})$ in Algorithm 2.

We use the following biases in our experiments to sample from $\Pi^{H_{\text{bias}}}$,

1. Speed Asymmetry. Humans and agents do not have the same speed of action and decision-making, and this speed varies between humans. We model this by taking an action from a trained SP model with probability p and taking no action otherwise. This improves training efficiency producing a population of agents with different behavior with a single SP model.
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.[2]

Algorithm 1 Learn SP human behavior prior

Input: Set P with Markov games representing different behavior priors.
Initialize $\Pi^{H_{\text{bias}}}$ to $\emptyset$.
for $m \in P$ **do**
  Train self-play policy, π^{sp}, for Markov game m.
  Add π^{sp} to $\Pi^{H_{\text{bias}}}$.
end for
Output: $\Pi^{H_{\text{bias}}}$.

Algorithm 2 Best-response to human prior policies

Input: $\Pi^{H_{\text{bias}}}$.
Initialize BR agent, $\pi^{\text{BR(H)}}$.
while $\pi^{\text{BR(H)}}$ not converged **do**
  Form minibatch from $\pi^{\text{BR(H)}}$ paired with elements of $\Pi^{H_{\text{bias}}}$.
  Use minibatch to update $\pi^{\text{BR(H)}}$.
end while
Output: $\pi^{\text{BR(H)}}$.

5 Experiments

Overcooked. We utilize the Overcooked environment introduced by [4] due to its combination of strategy and motion coordination challenges. 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 only the Cramped Room layout. The primary challenge lies in the agents' ability to navigate the environment, interact with objects, and coordinate their strategies. The agent can take six actions: `up`, `down`, `right`, `left`, `noop`, and `interact`. For training RL agents, we use proximal policy optimization [16] implemented in the JaxMARL library [14] using the same state encoding and network architecture.

Results. Our results in Table 1 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 BR(k) agent is trained as best response to a set of k SP agents, similar to [19]. Our method, BR (H_{Speed}), is a best-response to SP($p = 1$) and SP ($p = 0$), where p is the probability of the agent taking a `noop` action. The results show us that the SP model has the lowest return, and increasing the number of SP agents in the BR increase the return. This is expected as the BR agent with more

[2]This experiment is not included in this preliminary work but will be included in the future.

Method	Self-Play	BR(8)	BR(16)	BR(32)	**BR (H_{Speed})**
Avg. Episodic Returns	91.2 ± 3.1	95.6 ± 2.2	100.8 ± 2.5	105.0 ± 2.0	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 $p = 1, 0$.

SP agents is able to adapt to more partner behaviors. We also see that our approach, using only two SP agents for the BR, is able to significantly outperform even BR($k = 32$), validating the increased efficiency due to the included bias of variable agent reaction speed.

6 Conclusion and Future Work

This research explores an approach of incorporating well-studied systematic biases in human behavior to enhance reinforcement learning (RL) systems for fully cooperative games. By modifying the Markov game framework to create biased RL agents and subsequently training a best-response agent to interact with humans, we aim to develop solutions that can adapt well to human behavior without the need for task-specific human data.

Our preliminary results indicate that even simple behavioral biases can lead to significant improvements in learning efficiency. However, this work is still in progress and requires further experimentation to validate these findings comprehensively. Future work will focus on implementing a broader array of cognitive biases and conducting user experiments to evaluate their effectiveness.

Our approach exemplifies how human biases can be integrated into reinforcement learning systems within a cooperative framework. An important avenue for future research is to determine which biases are beneficial in different domains and how these biases can be systematically translated into objectives for learning agents. We hope our work contributes to a deeper understanding of how human behavioral biases can be harnessed to improve AI systems across diverse applications.

References

[1] G. Ainslie and N. Haslam. Hyperbolic discounting. In G. Loewenstein and J. Elster, editors, *Choice over time*. Russell Sage Foundation, 1992.

[2] J. R. Anderson. *The adaptive character of thought*. Psychology Press, 1990.

[3] S. Bansal, J. Xu, A. Howard, and C. Isbell. Planning for human-robot parallel play via bayesian nash equilibrium inference. In *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[4] M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems*, pages 5175–5186, 2019.

[5] G. Gigerenzer and D. G. Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996.

[6] T. L. Griffiths. Understanding human intelligence via human limitations. *Trends in Cognitive Sciences*, 24(11):873–883, 2020.

[7] J. C. Harsanyi and R. Selten. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

[8] H. Hu, A. Lerer, A. Peysakhovich, and J. Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

[9] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[10] F. Lieder and T. L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1: 1–60, 2020.

[11] A. Lupu, B. Cui, H. Hu, and J. Foerster. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*, pages 7204–7213. PMLR, 2021.

[12] R. Mirsky, I. Carlucho, A. Rahman, E. Fosong, W. Macke, M. Sridharan, P. Stone, and S. V. Albrecht. A survey of ad hoc teamwork research. In D. Baumeister and J. Rothe, editors, *Multi-Agent Systems*, pages 275–293, Cham, 2022. Springer International Publishing.

[13] M. I. Norton, D. Mochon, and D. Ariely. The ikea effect: When labor leads to love. *Journal of consumer psychology*, 22(3):453–460, 2012.

[14] A. Rutherford, B. Ellis, M. Gallici, J. Cook, A. Lupu, G. Ingvarsson, T. Willi, A. Khan, C. Schroeder de Witt, A. 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*, pages 2444–2446, 2024.

[15] T. C. Schelling. *The Strategy of Conflict*. Harvard university press, 1980.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[17] H. A. Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.

[18] P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein. Ad hoc autonomous agent teams: collaboration without pre-coordination. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, page 1504–1509. AAAI Press, 2010.

[19] D. Strouse, K. McKee, M. Botvinick, E. Hughes, and R. Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515, 2021.

[20] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[21] J. Von Neumann and O. Morgenstern. *The theory of games and economic behavior*. Princeton University Press, 1944.

[22] M. Yang, M. Carroll, and A. Dragan. Optimal behavior prior: Data-efficient human models for improved human-ai collaboration. In *Human in the Loop Learning (HiLL) Workshop at NeurIPS*, 2022.

[23] C. Yu, J. Gao, W. Liu, B. Xu, H. Tang, J. Yang, Y. Wang, and Y. Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

[24] R. Zhao, J. Song, Y. Yuan, H. Hu, Y. Gao, Y. Wu, Z. Sun, and W. Yang. Maximum entropy population-based training for zero-shot human-ai coordination. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5):6145–6153, Jun. 2023.