Reinforcement Learning with Cognitive Bias for Human-AI Ad Hoc Teamwork

Shray Bansal[1] **Jin Xu**[1] **Miguel Morales**[1] **Jonathan Streater**[2]
[1] *Georgia Institute of Technology* [2] *University of Florida*

Ayanna Howard[3] **Charles L. Isbell**[4]
[3] *The Ohio State University* [4] *University of Wisconsin-Madison*

Abstract

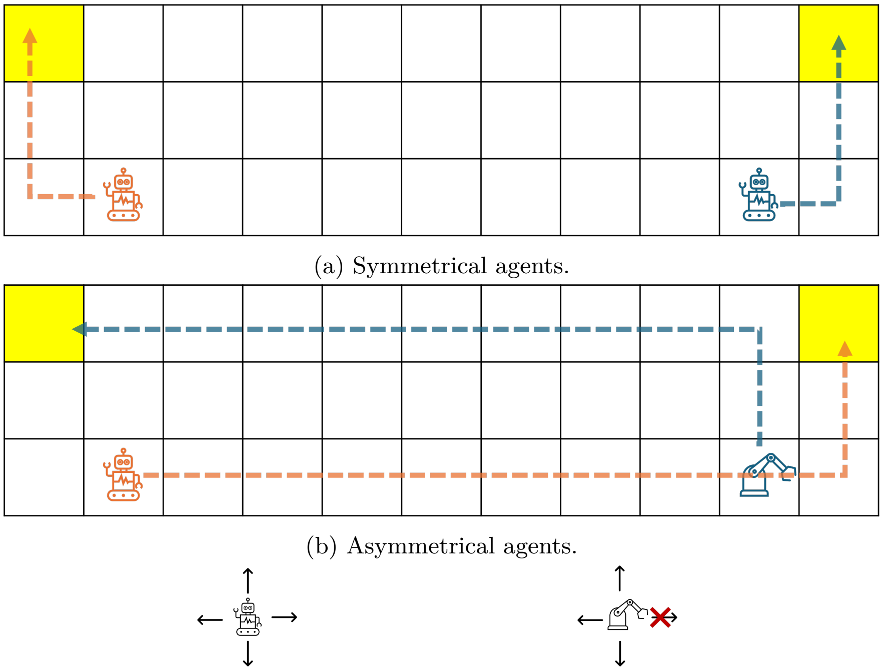
Recent advances in multiagent reinforcement learning have enabled artificial agents to coordinate effectively in complex domains; however, these agents can struggle to coordinate with humans, in part due to their implicit but inaccurate assumptions about optimal decision-making and behavioral homogeneity 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 perform that can coordinate with agents learned from human behavior 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 (Zhao et al., 2023; Lupu et al., 2021). Strouse et al. (2021) 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 (Zhao et al., 2023) in the objective to induce diversity in goal-driven behavior while reducing the number of agents sampled.

Corresponding author: Shray Bansal (sbansal34@gatech.edu)



(a) Symmetrical agents.

(b) Asymmetrical agents.

(c) Agent action spaces.

Figure 1: Example game where the goal of the agents is to occupy both 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 due to the inability one of the agents to move right as shown in (c).

In our problem, human-AI AHT, the interacting agents are known to be human and 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, we use 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 with other diversity metrics. 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 the grid-world coordination scenario in Fig. 1. In Fig. 1(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 Fig. 1(b), we assume that the agents are 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. Although, this is a simple example, we claim that incorporating simple biases from human behavior can have a large effect on coordination.

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 Mirsky et al. (2022) and the related problem of zero-shot coordination (ZSC) Hu et al. (2020). Although these terms are sometimes used interchangeably, we adopt the distinction provided by Treutlein et al. (2021). According to their classification, a ZSC agent assumes that its interaction will be with agents that are optimized to coordinate well with ZSC agents, without having access to these partners during training. This makes ZSC a special case of the general AHT problem. Next, we review this problem and some prior solutions from the lens of game theory.

Game Theory. The equilibrium selection problem (Harsanyi & Selten, 1988) arises in pure coordination games with multiple equilibria because it is within the agents' best interests to coordinate on a single equilibrium, but this coordination is challenging without prior agreement. This problem is similar to AHT, where agents are also cooperative but may fail to coordinate due to a lack of prior interaction. Solutions to equilibrium selection can be categorized into two types: those that rely only on endogenous information about the game, and, those that also incorporate exogenous information about the agents. We use a similar classification to describe prior work in AHT below.

Examples using only endogenous information in AHT include: Hu et al. (2020) handling game symmetries in Hanabi by learning permutation-robust policies, Strouse et al. (2021) training agents as best-response to multiple Nash equilibria by learning multiple SP policies and pairing them with it, and, Zhao et al. (2023) include a maximum entropy objective to increase SP policy diversity paired with the best response agent.

In game theory, exogenous information based on the the agent's options independent of the rewards of the game has been used to explain human behavior. For instance, Schelling (1980) showed that humans were able to coordinate significantly better than chance when playing a simple coordination game where agents aim to choose the same side of a coin. In human-AI coordination examples include: Bansal et al. (2020) using information about human behavior and social norms to group Nash equilibria and adapt online to human behavior in a table-top manipulation task, and, Yu et al. (2023) leveraging information about human bias to generate multiple event-based reward functions and learn a BR policy in Overcooked. Our approach also utilizes exogenous information in the form of systematic biases in human behavior. However, unlike Yu et al. (2023), which sample reward functions based on defined game *events*, we sample policies based on human behavioral traits without introducing a new reward structure.

Cognitive Bias: Starting soon after scientists began formalizing and describing human behavior as rational actors (Von Neumann & Morgenstern (1944)), they also began observing and describing systematic human deviations from these classic notions of rationality via logic, probability theory, and, expected utility (Tversky & Kahneman (1974); Kahneman & Tversky (1979)). Known as cognitive biases, these systematic deviations, have been identified in myriad environments and contexts, challenging the idealized concept of humans as rational. They have been used to introduce frameworks like bounded rationality, bounded optimality, ecological rationality, rational analysis, and more recently, resource-rational analysis, that can account for them (Simon (1955; 1956); Gigerenzer & Goldstein (1996); Anderson (1990); Lieder & Griffiths (2020)). Their core idea is to describe and explain these cognitive agents functioning within environments and contexts that take the best advantage of limited cognitive resources. Understanding human cognition as optimal and general under limitations of time, computation and communication (Griffiths (2020)), 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, our research takes into account two human limitations: (1) limitation on human reaction speed to situational changes, and, (2) preference for immediate over future rewards (Ainslie & Haslam, 1992).

Availability of Data. When human behavior data is available, it can help to learn agents that successfully coordinate with humans, as shown by Carroll et al. (2019) in Overcooked. However, collecting this task-specific data for every scenario that AI agents will interact with humans is impractical. Also, human behavior can evolve over time, and, can vary in interactions with AI versus other humans. We aim to identify task-invariant properties of human behavior that are broadly applicable across various 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 Yang et al. (2022).

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, as defined by Stone et al. (2010), 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". 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, $\Pi_K = \{\pi^1, \dots, \pi^K\}$ as,

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

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, $\pi^{BR(H)} \in \mathcal{B}(\Pi^{H*})$. 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 approximate human 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 aim to improve *ad hoc* teamwork with unseen human teammates and reduce training time without requiring task-specific human training data by utilizing task-invariant properties of human behavior.

We define the term, behavior prior, to mean both skill asymmetry and cognitive bias. We use the following behavior priors 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, also this speed varies among humans. We model this by taking an action from a trained SP model with probability p and taking no action otherwise. This has the added benefit of improving training efficiency since a population of agents with differing behavior can be produced, by varying p, while using only a single trained 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.[1]

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, $H^{sp}$, for Markov game $m$.
    Add $H^{sp}$ 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 Carroll et al. (2019) 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 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), using the same state encoding provided and network architecture developed by them. We use proximal policy optimization (Schulman et al., 2017) to train these models.

Results. 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 BR(N_{SP}) agent is trained as best response to N_{SP} SP agents, similar to the approach used by Strouse et al. (2021). Our method, BR (H_{Speed}), is a

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

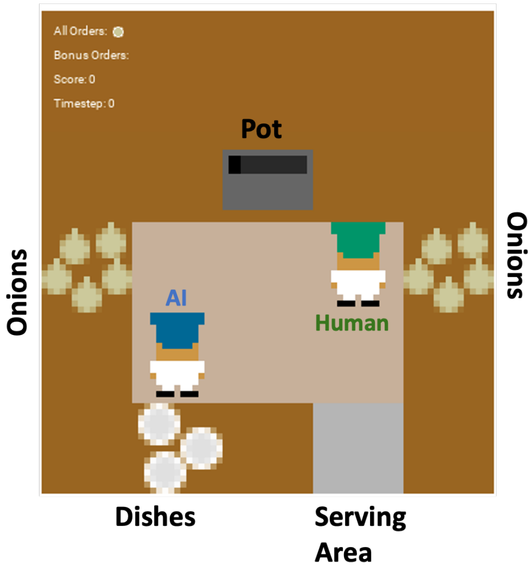


Figure 2: A human and an AI agent interact in a shared environment in the fully-cooperative game of Overcooked.

Method	Self-Play	BR(8)	BR(16)	BR(32)	**BR ($H_{\textbf{Speed}}$)**
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 $p = 1, 0$.

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 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($N_{SP} = 32$), validating the increased efficiency due to the included bias of variable agent reaction speed.

6 Conclusion and Future Work

This research explores an innovative approach by 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

George Ainslie and Nick Haslam. Hyperbolic discounting. 1992.

John R Anderson. *The adaptive character of thought*. Psychology Press, 1990.

Shray Bansal, Jin Xu, Ayanna Howard, and Charles Isbell. Planning for human-robot parallel play via bayesian nash equilibrium inference. In *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020. doi: 10.15607/RSS.2020.XVI.042.

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.

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996. doi: 10.1037/0033-295X.103.4.650.

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

John C Harsanyi and Reinhard Selten. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pp. 4399–4410. PMLR, 2020.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. doi: 10.2307/1914185.

Falk Lieder and Thomas 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. doi: 10.1017/S0140525X1900061X.

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.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. A survey of ad hoc teamwork research. In Dorothea Baumeister and Jörg Rothe (eds.), *Multi-Agent Systems*, pp. 275–293, Cham, 2022. Springer International Publishing. ISBN 978-3-031-20614-6.

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.

Thomas C Schelling. *The Strategy of Conflict*. Harvard university press, 1980.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pp. 99–118, 1955.

Herbert A Simon. Rational choice and the structure of the environment. *Psychological Review*, 63 (2):129–138, 1956. doi: 10.1037/h0042769.

Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1504–1509, 2010.

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.

Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism, method and open issues for zero-shot coordination. In *International Conference on Machine Learning*, pp. 10413–10423. PMLR, 2021.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124.

John Von Neumann and Oskar Morgenstern. *The theory of games and economic behavior*. Princeton University Press, 1944.

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

Chao Yu, Jiaxuan Gao, Weilin Liu, Botian Xu, Hao Tang, Jiaqi Yang, Yu Wang, and Yi Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

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.