Collaborating with Humans without Human Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Collaborating with humans requires rapidly adapting to their individual strengths, weaknesses, and preferences. Unfortunately, most standard multi-agent reinforcement learning techniques, such as self-play (SP) or population play (PP), produce agents that overfit to their training partners and do not generalize well to humans. Alternatively, researchers can collect human data, train a human model using behavioral cloning, and then use that model to train "human-aware" agents ("behavioral cloning play", or BCP). While such an approach can improve the generalization of agents to new human co-players, it involves the onerous and expensive step of collecting large amounts of human data first. Here, we study the problem of how to train agents that collaborate well with human partners without using human data. We argue that the crux of the problem is to produce a diverse set of training partners. Drawing inspiration from successful multi-agent approaches in competitive domains, we find that a surprisingly simple approach is highly effective. We train our agent partner as the best response to a population of self-play agents and their past checkpoints taken throughout training, a method we call Fictitious Co-Play (FCP). Our experiments focus on a two-player collaborative cooking simulator that has recently been proposed as a challenge problem for coordination with humans. We find that FCP agents score significantly higher than SP, PP, and BCP when paired with novel agent and human partners. Furthermore, humans also report a strong subjective preference to partnering with FCP agents over all baselines.

1 Introduction

Generating agents which collaborate with novel partners is a longstanding challenge for Artificial Intelligence (AI) [4, 34, 14, 48]. Achieving ad-hoc, zero-shot coordination [27, 61] is especially important in situations where an AI must generalize to novel human partners [5, 56]. Many successful approaches have employed human models, either constructed explicitly [13, 32, 49] or learnt implicitly [11, 55]. By contrast, recent work in competitive domains has shown that it is possible to reach human-level using model-free reinforcement learning (RL) without human data, via self-play [7, 8, 58, 59]. This begs the question: Can model-free RL without human data generate agents that can collaborate with novel humans?

We seek an answer to this question in the space of common-payoff games, where all agents work towards a shared goal and receive the same reward. Self-play (SP), in which an agent learns from repeated games played against copies of itself, does not produce agents that generalize well to novel co-players [9, 10, 18, 41]. Intuitively, this is because agents trained in self-play only ever need to coordinate with themselves, and so make for brittle and stubborn collaborators with new partners who act differently. Population play (PP) trains a population of agents, all of whom interact with each other [36]. While PP can produce agents capable of cooperation with humans in competitive team games [31], it still fails to produce robust partners for novel humans in pure common-payoff settings

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

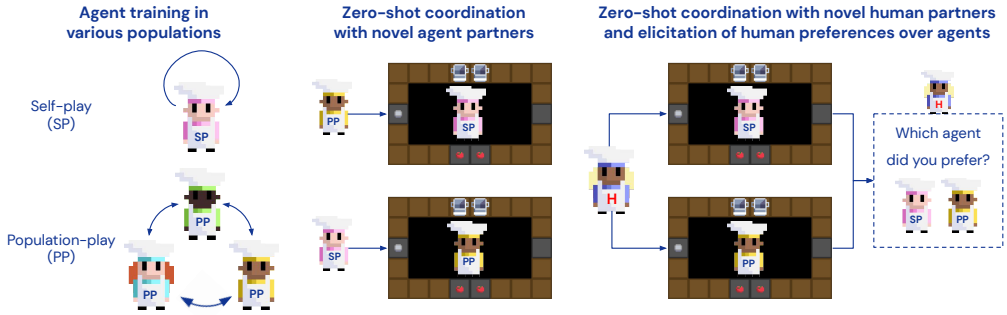


Figure 1: In this work, we evaluate a variety of agent training methods (Section 2) in zero-shot coordination with agents (Section 4). We then run a human-agent collaborative study designed to elicit human preferences over agents (Section 5).

[11]. PP in common-payoff settings naturally encourages agents to play the same way, reducing strategic diversity and producing agents not so different from self-play [21].

Our approach starts with the intuition that the key to producing robust agent collaborators is exposure to diverse training partners. We find that a surprisingly simple strategy is effective in generating sufficient diversity. We train N self-play agents varying only their random seed for neural network initialization. Periodically during training, we save agent "checkpoints" representing their strategy at that point in time. Then, we train an agent partner as the best-response to both the fully-trained agents and their past checkpoints. The different checkpoints simulate different skill levels, and the different random seeds simulate breaking symmetries in different ways. We refer to this agent training procedure as **Fictitious Co-Play (FCP)** for its relationship to fictitious self-play [6, 23, 24, 65].

We evaluate FCP in a fully-observable two-player common-payoff collaborative cooking simulator. Based on the game Overcooked [20], it has recently been proposed as a coordination challenge [11, 46, 68]. State-of-the-art performance in producing agents capable of generalization to novel humans was achieved in [11] via behavioral cloning (BC) of human data. More precisely, BC was used to produce models that can stand in as human proxies during training in simulation, a method we call behavioral cloning play (BCP). We demonstrate that FCP outperforms BCP in generalizing to both novel agent and human partners, and that humans express a significant preference for partnering with FCP over BCP. Our method avoids the cost and privacy concerns of collecting human data for training, while achieving better outcomes for humans at test time.

We summarize the novel contributions of this paper as follows:

1. We propose Fictitious Co-Play (FCP) to train agents capable of zero-shot coordination with humans (Section 2.1).
2. We demonstrate that FCP agents generalize better than SP, PP, and BCP in zero-shot coordination with a variety of held-out agents (Section 4.2).
3. We propose a rigorous human-agent interaction study with behavioral analysis and participant feedback (Section 5.1).
4. We demonstrate that FCP significantly outperforms the BCP state-of-the-art, both in task score and in human partner preference (Section 5.2).

2 Methods

2.1 Fictitious Co-Play (FCP)

Diverse training conditions have been shown to make agents more robust, from environmental variations [50] to heterogeneity in training partners [64]. We seek to train agents that are robust partners for humans in common-payoff games, and so extend this line of work to that setting.

One important challenge in collaborating with novel partners is dealing with symmetries [27]. For example, two agents A and B facing each other may move past each other by A going left and B going right, or vice versa. Both are valid solutions, but a good agent partner will adaptively switch between these conventions if a human clearly prefers one over the other. A second important challenge is

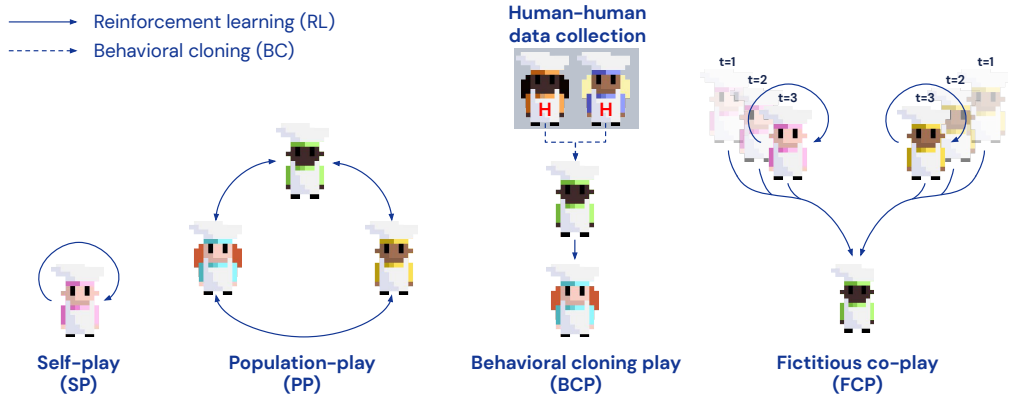


Figure 2: The four agent training methods we evaluate in this work. **Self-play (SP)** where an agent learns with itself, **population-play (PP)** where a population of agents are co-trained together, and **behavioral cloning play (BCP)** where data from human games is used to create a behaviorally cloned agent which an RL agent is then trained with. In our method, **Fictitious Co-Play (FCP)**, N self-play agents are trained independently and checkpointed throughout training. An agent is then trained to best respond to the entire population of SP agents and their checkpoints.

dealing with variations in skill level. Good agent partners should be able to assist both highly-skilled partners, as well as partners who are still learning.

Fictitious co-play (FCP) is a simple two-stage approach for training agents that overcomes both of these challenges (Figure 2, right). In the first stage, we train a pool of diverse training partners for an FCP agent. To allow the pool to represent different conventions for symmetry breaking, we train N partner agents in self-play. Since these partners are trained independently, they can arrive at different arbitrary conventions for breaking symmetries. To allow the pool to represent different skill levels, we use multiple checkpoints of each self-play partner throughout training. The final checkpoint represents a fully-trained "skillful" partner, while earlier checkpoints represent less skilled partners. Notably, by using multiple checkpoints per partner agent, this additional diversity in skill incurs no extra training cost.

In the second stage, we train an FCP agent as the best response to the pool of diverse partners created in the first stage. Importantly, the partner parameters are frozen and thus FCP must learn to adapt to partners, rather than expect partners to adapt to it. In this way, FCP agents are prepared to follow the lead of human partners, and learn a general policy across a range of strategies and skills. We call our method "fictitious" co-play for its relationship to fictitious self-play in which competitive agents are trained with past checkpoints (in that case, to avoid strategy cycling) [6, 23, 24, 36, 64].

2.2 Baselines and ablations

We compare FCP agents to the three baseline training methods listed below, each varying only in their set of training partners, with the RL algorithm and architecture consistent across all agents:

1. Self-play (SP), where agents learn solely through interaction with themselves.
2. Population-play (PP), where a population of agents are co-trained through random pairings.
3. Behavioral cloning play (BCP), where an agent is trained with a BC model of a human [11].

We also evaluate three variations on FCP to better understand the conditions for its success:

1. To test the importance of including past checkpoints in training, we evaluate an ablation of FCP in which agents are trained only with the converged checkpoints of their partners (FCP_{-T} for "FCP minus time").
2. To test whether FCP would benefit from additional diversity in its partner population, we evaluate an augmentation of FCP in which the population of SP partners varies not just in random seed, but also in architecture (FCP_{+A} for "FCP plus architectural variation").
3. To test whether architectural variation can serve as a full replacement for playing with past checkpoints, we evaluate the combination of both modifications ($\text{FCP}_{-T,+A}$).

2.3 Environment

Following prior work on zero-shot coordination in human-agent interaction, we study the Overcooked environment (see Figure 3) [11, 12, 35, 46, 66]. We draw particular inspiration from the environment in Carroll et al. [11]. For full details, see Appendix A.

In this environment, players are placed into a gridworld kitchen as chefs and tasked with delivering as many cooked dishes of tomato soup as possible within an episode. This involves a series of sequential high-level actions which both players can contribute to: collecting tomatoes, depositing them into cooking pots, letting the tomatoes cook into soup, collecting a dish, getting the soup, and delivering it. Upon a successful delivery, both players are rewarded equally.

To effectively complete the task, players must learn to navigate the kitchen and interact with objects in the correct order, all while maintaining awareness of their partner's behavior to coordinate with them. This environment therefore presents the challenges of both movement and strategic coordination.

Each player observes an egocentric RGB view of the world, and at every step can perform one of six actions: `stand still`, `move {up, down, left, right}`, `interact`. The behavior of `interact` varies based on the cell which the player is facing (e.g. place tomato on counter).

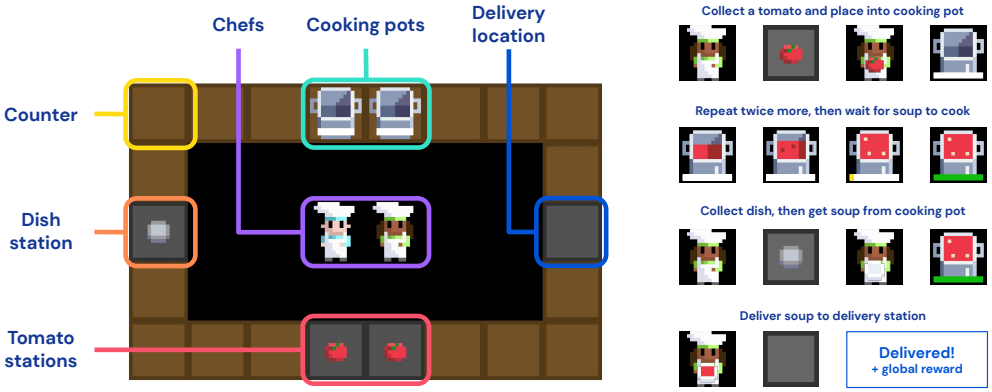


Figure 3: **The Overcooked environment:** a two-player common-payoff game in which players must coordinate to cook and deliver soup.

Figure 4: **Layouts:** the kitchens which agents and humans play in, each emphasizing different coordination strategies. Highlighted in bold are the terms used to refer to each in the rest of this paper.

2.4 Implementation details

Here we highlight several key implementation details for our training methods. For full details, including the architectures, hyperparameters, and compute used, please see Appendix B.

For our reinforcement learning agents, we use the V-MPO [60] algorithm along with a ResNet [22] plus LSTM [25] architecture which we found led to optimal behavior across all layouts. Agents are trained using a distributed set of environments running in parallel [15], each sampling two agents from the training population to play together every episode.

Both PP and FCP are trained with a population size of $N = 32$ agents which are sampled uniformly. For FCP, we use 3 checkpoints for each agent, therefore incurring no additional training burden: (1) at initialization (i.e. a low-skilled agent), (2) at the end of training (i.e. a fully-trained expert agent), and (3) at the middle of training, defined as when the agent reaches 50% of its final reward (i.e. an average-skilled agent). When varying architecture for the training partners of the FCP_{+A} and

FCP$_{-T,+A}$ variants, we vary whether the partners use memory (i.e. LSTM vs not) and the width of their policy and value networks (i.e. 16 vs 256). In total, we train 8 agents for each of the 4 combinations, leaving the total population size of $N = 32$ unchanged, ensuring a fair comparison.

To train agents via behavioral cloning [53], we use Acme's [26] open-source implementation to learn a policy from human gameplay data. Specifically, we collected 5 human-human trajectories of length 1200 time steps for each of the 5 layouts, resulting in 60k total environment steps. We divide this data in half and train two BC agents: (1) a partner for training a BCP agent, and (2) a "human proxy" partner for agent-agent evaluation. Similar to Carroll et al. [11], we use a set of feature-based observations for the agents (as opposed to RGB) with comparable results – performing well on 3 layouts (`asymmetric`, `cramped`, and `ring`) but poorer on the other 2 (`circuit` and `forced`).

3 Related work

3.1 Ad-hoc team play

There is a large and diverse body of literature on ad-hoc team-play [42, 61], also known as zero-shot coordination [27]. Prior work based in game-theoretic settings has suggested the benefits of planning [67], online learning [47], and novel solution concepts [2], to name a few examples. More recently, multi-agent deep reinforcement learning has provided the tools to scale to more complex gridworld or continuous control settings, leading to work on hierarchical social planning [33], adapting to existing social conventions [37, 57], trajectory diversity [42], and theory of mind [13]. Ad-hoc team-play among novel agent partners is also an object of active study in the emergent communication literature [9, 10, 40]. This prior work has tended to focus on generalization to held-out agent partners as a proxy for human co-players.

Collaborative play with novel humans has been evaluated more actively in the context of training agent assistants; see for instance [52, 62]. To our knowledge, our FCP agents represent the state-of-the-art in coordinating with novel human partners on an equal footing of capabilities in a rich gridworld environment, as measured by the challenge tasks in Carroll et al. [11].

3.2 Diversity in multi-agent reinforcement learning

In multi-agent reinforcement learning, agents that train with behaviorally diverse populations of game partners tend to demonstrate stronger performance than their self-play counterparts. For example, across a range of multi-agent games, generalization to held-out populations can be improved by training larger and more diverse populations [12, 39, 46]. In mixed-motive settings, cooperation among agents can be encouraged through social diversity, such as in player preferences and rewards [3, 44]. Similarly, competitiveness can be optimized through selective matchmaking between increasingly diverse agents [21, 36, 65].

Despite the increased focus on improving multi-agent performance, evaluation has typically been constrained to agent-agent settings; high-performing agents have infrequently been evaluated with humans, particularly in non-competitive domains [14]. We expand this growing literature, showing that training with diversity is a powerful approach for effective human-agent collaboration.

3.3 Human-agent interaction

In recent years, increased attention has been directed to designing machine learning agents capable of collaborating with humans [38, 52, 63, 69] (see also [14] for a wider review on Cooperative AI).

A key preceding entry in this research area is Carroll et al. [11], who similarly investigated human-agent coordination in the Overcooked environment. We use their method (BCP) as a baseline throughout our agent-agent and human-agent experiments (Section 2.2). Relative to BCP, our approach removes the need for the expensive step of human data collection for agent training. Furthermore, through our novel human-agent experimental design, we go beyond objective performance metrics to compare the subjective preferences that agents generate. For a detailed comparison of methods and results, see Appendix E.

4 Zero-shot coordination with agents

In this section, we evaluate our FCP agent, its ablations, and the baselines with held-out agents.

4.1 Evaluation method: collaborative evaluation with agent partners

Our primary concern in this work is generalization to novel *human* partners (as investigated in Section 5). However, just as collecting human-human data for behavioral cloning is expensive, so too is evaluating agents with humans. Consequently, we instead use generalization to held-out *agent* partners as a cheap proxy of performance with humans. This is then used to guide our model selection process, allowing us to be more targeted with the agents we select for our human-agent evaluations.

We evaluate with three held-out populations,

1. A BC model trained on human data, H_{proxy}, intended as a proxy of generalization to humans as done by Carroll et al. [11].
2. A set of self-play agents varying in seed, architecture, and training time. These are intended to test generalization to a diverse yet still skillful population.
3. Randomly initialized agents intended to test generalization to low-skill partners.

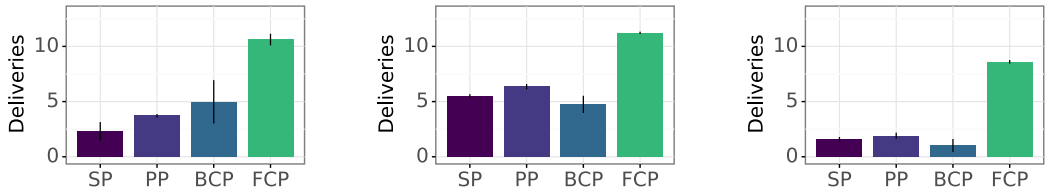
For all of our results, we report the average number of deliveries made by both players within an episode, aggregated across the 5 different layouts from Figure 4 (with the per-layout results reported in the Appendix). We estimate mean and standard error across 5 random seeds. For each seed, we evaluate the agent with all members of the held-out population for 10 episodes per agent-partner pair.

4.2 Results

Finding 1: FCP significantly outperforms all baselines

To begin, we compare our FCP agent and the baselines when partnered with the three held-out populations introduced above. As can be seen in Figure 5, FCP significantly outperforms all baselines when partnered with all three held-out populations. Notably, it performs better than BCP with H_{proxy}, even though BCP trains with such a model and FCP does not. Similar to Carroll et al. [11], we find that BCP significantly outscores SP.

When paired with a randomly initialized partner which behaves suboptimally, we see an even greater difference between FCP and the baselines. Given that FCP is trained with non-held-out versions of such agents, it may not be surprising that it does so well with partners that behave poorly. However, what is surprising is how brittle the other training methods are. This suggests that they may not perform well with humans who are not highly skilled players, which we will see in Section 5.



(a) With H_{proxy}. (b) With diverse SP agents. (c) With random agents.

Figure 5: **Agent-agent collaborative evaluation:** Performance of each agent when partnered with each of the held-out populations in episodes of length $T = 540$. Importantly, FCP scores higher than all baselines with a variety of test partners. Error bars represent standard deviation over five random training seeds. Plots aggregate data across kitchen layouts; results calculated by individual layout can be found in the Appendix.

Finding 2: Training with past checkpoints is the most beneficial variation for performance

Next, we investigate how the different training partner variations influence FCP's performance. In particular, we separately ablate the past checkpoints (T) and architecture (A) variations, evaluating them with the same partners as in Figure 5. The results of this evaluation are presented in Table 1.

Partner	FCP	FCP$_{-T}$	FCP$_{+A}$	FCP$_{-T,+A}$
H_{proxy}	10.6 ± 0.5	4.7 ± 0.4	9.9 ± 0.6	7.0 ± 0.8
Diverse SP	11.2 ± 0.1	6.9 ± 0.1	11.1 ± 0.4	8.6 ± 0.4
Random	8.6 ± 0.2	1.0 ± 0.1	8.4 ± 0.4	3.2 ± 0.5

Table 1: **Ablation results:** Performance of each variation of FCP – training with past partner checkpoints (T for time) and adding partner variation in architecture (A). Scores are mean deliveries with standard deviation over 5 random seeds. Notably, we find that the inclusion of past checkpoints is essential for strong performance (FCP > FCP$_{-T}$), and additionally including architectural variation does not improve performance (FCP $\approx$ FCP$_{+A}$). However, architectural variation is better than no variation, improving performance when past checkpoints are not available (FCP$_{-T,+A}$ > FCP$_{-T}$).

Comparing the FCP and FCP$_{-T}$ columns, we see that removing past checkpoints from training significantly reduces performance. Comparing the FCP and FCP$_{+A}$ columns, we see that adding architectural variation to the training population offers no improvement over training with past checkpoints. However, comparing the FCP$_{-T}$ and FCP$_{-T,+A}$ columns, we see that without training with past checkpoints, architectural variation in the population does improve performance.

5 Zero-shot coordination with humans

Ultimately, our goal is to develop agents capable of coordinating with novel human partners. In this section, we run an online study to evaluate our FCP agent and the baseline agents in collaborative play with human partners.

Figure 6: **Human-agent collaborative study:** For our human-agent collaboration study, we recruited participants online to play games with FCP and baseline agents. Participants played a randomized sequence of episodes with different agent partners and kitchen layouts. After every two rounds, participants reported the direction and strength of their preference between their agent partners from those two episodes.

5.1 Evaluation method: collaborative evaluation with human participants

To test how effectively FCP's performance generalizes to human partners, we recruited participants from Prolific [51] for an online collaboration study ($N = 114$; 37.7% female, 59.6% male, 1.8% nonbinary; median age between 25–34 years). We used a within-participant design for the study: each participant played with a full cohort of agents (i.e., generated through every training method). This design allows us to evaluate the effects of training method on objective performance as well as subjective preferences.

Participants first read game instructions and played a short tutorial episode guiding them through the dish preparation sequence (see Appendix for instruction text and study screenshots). Participants then played 20 episodes with a randomized sequence of agent partners and kitchen layouts. Episodes lasted $T = 300$ steps (1 minute) each. After every two episodes, participants reported their preference over the agent partners from those episodes on a five-point Likert-type scale. After playing all 20 episodes, participants completed a debrief questionnaire collecting standard demographic information and open-ended feedback on the study. Appendix contains full details of our study design, including independent ethical review.

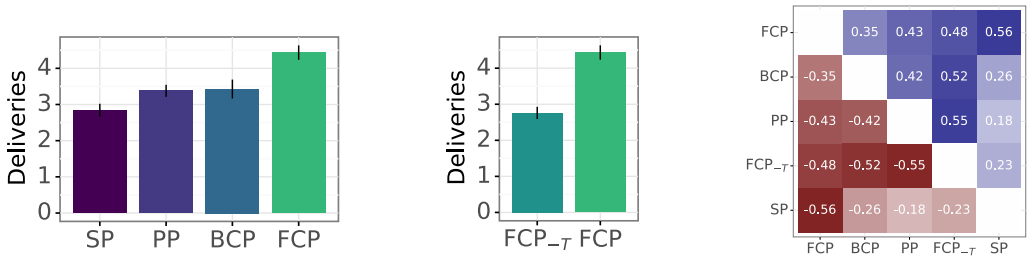
5.2 Results

Finding 1: FCP coordinates best with humans, achieving the highest scores across all maps

To begin, we compare the objective team performance supported by our FCP agent and the baseline agents. The strong FCP performance observed in agent-agent play effectively generalizes to human-agent collaboration: the FCP-human teams significantly outperform all other agent-human teams (Figure 7a). Echoing the results from our agent-agent ablation experiments (Table 1), the inclusion of past checkpoints in training proves critical for FCP's strong performance. Teams composed of human and FCP_{-T} agents achieve significantly lower scores than human-FCP teams (Figure 7b). Similar to Carroll et al. [11], we find that BCP outscores SP when collaborating with human players.

Finding 2: Participants prefer FCP over all baselines

FCP's strong collaborative performance carries over to our participants' subjective partner preferences. Participants expressed a significant preference for FCP partners over all other agents, including BCP (Figure 7c). Notably, though human-BCP and human-PP teams scored comparably in terms of deliveries, participants reported significantly preferring BCP over PP.



(a) Number of deliveries by partner (FCP and baselines).

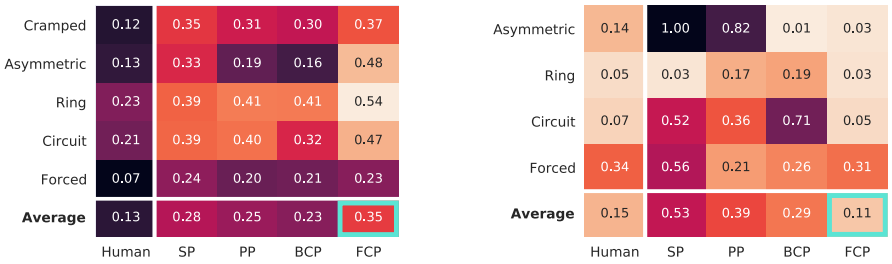
(b) Number of deliveries by partner (FCP_{-T} and FCP).

(c) Participant preference for row partner over column partner.

Figure 7: **Human-agent collaborative evaluation**: Evaluation and preference metrics from human-agent play in episodes of length $T = 300$. Error bars represents 95% confidence intervals, calculated over episodes. Plots aggregate data across kitchen layouts; results calculated by individual layout can be found in the Appendix.

5.3 Exploratory behavioral analysis

To better understand how the human-agent scores and preferences may have arisen, here we analyze the resulting action trajectories of each human and agent player in our experiment.



(a) Proportion of episode spent moving.

(b) Differences in pot preference.

Figure 8: **Behavioral analysis:** (a) FCP is able to move most frequently (35% of the time), corresponding to the best movement coordination with human partners. (b) FCP exhibits the most equal preferences over cooking pots (0.11 difference), aligning with human preferences. Values are calculated as the absolute difference in preferences between the two pots; 1 indicates that the player only uses one of the two available pots, while 0 indicates that the player uses both pots equally.

Finding 1: FCP exhibits the best movement coordination with humans

First, we investigate how much each player moves in an episode (Figure 8a), where moving in a higher fraction of timesteps may suggest fewer collisions and thus better coordination with a partner. Notably, we observe two results: (1) humans rarely move, a behavior which is out-of-distribution for typical training methods (e.g. SP, PP) but is seen in the training distribution for BCP and FCP. (2) FCP moves the most on all layouts other than Forced, suggesting it is better at coordinating its movement strategy with its partner. This result was also reported by human participants, for example: "I noticed that some of my partners seemed to know they needed to move around me, while others seemed to get 'stuck' until I moved out of their way" (see Appendix D for more examples).

Finding 2: FCP's preferences over cooking pots aligns best with that of humans

Next, we investigate whether there was a preference for a specific cooking pot in the layouts which included two cooking pots (Figure 8b). To do this, we calculate the difference in the number of times each pot was used by each player, where a high value indicates a strong preference for one pot and a low value indicates more equal preference for the two pots.

As can be seen in the FCP column, our agent typically has the most aligned preferences with that of humans (0.11 for FCP to 0.14 for humans). Behaviorly speaking, this means that our agent prefers one cooking pot over the other 55.5% of the time (i.e. a 0.11 point difference). In contrast, all other agents have a strong preference for a single pot. This is a non-adaptive strategy which generalizes poorly to typical human behavior of using both pots, leading to worse performance.

6 Discussion

Summary. In this work, we investigated the challenging problem of zero-shot collaboration with humans without using human data in the training pipeline. To accomplish this, we introduced Fictitious Co-Play (FCP) – a surprisingly simple yet effective method based on creating a diverse set of training partners. We found that FCP agents scored significantly higher than all baselines when partnered with both novel agent and human partners. Furthermore, through a rigorous human-agent experimental design, we also found that humans reported a strong subjective preference to partnering with FCP agents over all baselines.

Limitations and future work. Our method currently relies on the manual process of initially training and selecting a diverse set of partners. Not only is this time consuming, it is also prone to researcher biases which may negatively influence the behavior of the created agents. To address this, methods for automatically generating partner diversity for common-payoff games may be important. For example through adaptive population matchmaking as has been done in competitive zero-sum games [64], or through auxiliary objectives that explicitly encourage behavioral diversity [16, 42, 43].

Our method requires a known and fixed reward function. We also focus on one domain in order to compare with prior work which has argued that human-in-the-loop training is necessary. Consequently, the resulting agents are only designed to adaptively collaborate on a single task, and not to infer human preferences in general [1, 29, 54]. Before such methods can be deployed successfully, more domains and tasks should be studied to ensure these methods generalize safely.

Societal impact. A challenge for this line of work is ensuring agent values are aligned with human values (i.e. the AI value alignment problem [19, 54]). Our method has no guarantees that the resulting agent's values align with its potential partners. This could therefore lead to negative societal impacts if applied incorrectly or optimized for harmful metrics. Furthermore, methods like ours may produce exploitable agents which exacerbate existing inequalities and biases.

One strategy for mitigating these risks is the use of human preference data. Such data could be used to fine-tune and filter trained agents before deployment, encouraging better alignment with human values. Relatedly, targeted research on human beliefs and impressions of AI [45], as well as how they interact, would help inform agent design for positive societal impact. For instance, one could include specific biases that reinforce fairness and equity [17, 28].

Conclusion. We proposed a method which is both effective at collaborating with humans and simple to implement. We also presented a rigorous and general methodology for evaluating with humans and eliciting their preferences. Together, these establish a strong foundation for future research on the important challenge of human-agent collaboration for benefiting society.

References

[1] J. Abramson, A. Ahuja, I. Barr, A. Brussee, F. Carnevale, M. Cassin, R. Chhaparia, S. Clark, B. Damoc, A. Dudzik, et al. Imitating interactive intelligence. *arXiv preprint arXiv:2012.05672*, 2020.

[2] S. V. Albrecht and S. Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. *CoRR*, abs/1506.01170, 2015. URL http://arxiv.org/abs/1506.01170.

[3] B. Baker. Emergent reciprocity and team formation from randomized uncertain social preferences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b63c87b0a41016ad29313f0d7393cee8-Abstract.html.

[4] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, I. Dunning, S. Mourad, H. Larochelle, M. G. Bellemare, and M. Bowling. The hanabi challenge: A new frontier for AI research. *CoRR*, abs/1902.00506, 2019. URL http://arxiv.org/abs/1902.00506.

[5] A. Bauer, D. Wollherr, and M. Buss. Human–robot collaboration: a survey. *International Journal of Humanoid Robotics*, 5(01):47–66, 2008.

[6] G. W. Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

[7] N. Brown and T. Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. ISSN 0036-8075. doi: 10.1126/science.aao1733. URL https://science.sciencemag.org/content/359/6374/418.

[8] N. Brown and T. Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019. ISSN 0036-8075. doi: 10.1126/science.aay2400. URL https://science.sciencemag.org/content/365/6456/885.

[9] K. Bullard, F. Meier, D. Kiela, J. Pineau, and J. N. Foerster. Exploring zero-shot emergent communication in embodied multi-agent populations. *CoRR*, abs/2010.15896, 2020. URL https://arxiv.org/abs/2010.15896.

[10] K. Bullard, D. Kiela, J. Pineau, and J. N. Foerster. Quasi-equivalence discovery for zero-shot emergent communication. *CoRR*, abs/2103.08067, 2021. URL https://arxiv.org/abs/2103.08067.

[11] 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.

[12] R. Charakorn, P. Manoonpong, and N. Dilokthanakul. Investigating partner diversification methods in cooperative multi-agent deep reinforcement learning. In *International Conference on Neural Information Processing*, pages 395–402. Springer, 2020.

[13] R. Choudhury, G. Swamy, D. Hadfield-Menell, and A. D. Dragan. On the utility of model learning in hri. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 317–325. IEEE, 2019.

[14] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. Open problems in cooperative ai, 2020.

[15] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.

[16] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function, 2018.

[17] E. Fehr and K. M. Schmidt. A Theory of Fairness, Competition, and Cooperation*. *The Quarterly Journal of Economics*, 114(3):817–868, 08 1999. ISSN 0033-5533. doi: 10.1162/003355399556151. URL https://doi.org/10.1162/003355399556151.

[18] J. Foerster, F. Song, E. Hughes, N. Burch, I. Dunning, S. Whiteson, M. Botvinick, and M. Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1942–1951. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/foerster19a.html.

[19] I. Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

[20] G. T. Games. Overcooked, 2016. URL https://store.steampowered.com/app/448510/Overcooked/.

[21] M. Garnelo, W. M. Czarnecki, S. Liu, D. Tirumala, J. Oh, G. Gidel, H. van Hasselt, and D. Balduzzi. Pick your battles: Interaction graphs as population-level objectives for strategic diversity. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 1501–1503, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] J. Heinrich and D. Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

[24] J. Heinrich, M. Lanctot, and D. Silver. Fictitious self-play in extensive-form games. In *International conference on machine learning*, pages 805–813. PMLR, 2015.

[25] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[26] M. Hoffman, B. Shahriari, J. Aslanides, G. Barth-Maron, F. Behbahani, T. Norman, A. Abdolmaleki, A. Cassirer, F. Yang, K. Baumli, S. Henderson, A. Novikov, S. G. Colmenarejo, S. Cabi, C. Gulcehre, T. L. Paine, A. Cowie, Z. Wang, B. Piot, and N. de Freitas. Acme: A research framework for distributed reinforcement learning. *arXiv preprint arXiv:2006.00979*, 2020. URL https://arxiv.org/abs/2006.00979.

[27] 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.

[28] E. Hughes, J. Z. Leibo, M. G. Phillips, K. Tuyls, E. A. Duéñez-Guzmán, A. G. Castañeda, I. Dunning, T. Zhu, K. R. McKee, R. Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *arXiv preprint arXiv:1803.08884*, 2018.

[29] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei. Reward learning from human preferences and demonstrations in atari. *CoRR*, abs/1811.06521, 2018. URL http://arxiv.org/abs/1811.06521.

[30] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

[31] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castañeda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, N. Sonnerat, T. Green, L. Deason, J. Z. Leibo, D. Silver, D. Hassabis, K. Kavukcuoglu, and T. Graepel. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443): 859–865, 2019. ISSN 0036-8075. doi: 10.1126/science.aau6249. URL https://science.sciencemag.org/content/364/6443/859.

[32] S. Javdani, S. S. Srinivasa, and J. A. Bagnell. Shared autonomy via hindsight optimization. *Robotics science and systems: online proceedings*, 2015, 2015.

[33] M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil, M. L. Littman, and J. B. Tenenbaum. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *CogSci*, 2016.

[34] G. Klien, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19 (6):91–95, 2004.

[35] P. Knott, M. Carroll, S. Devlin, K. Ciosek, K. Hofmann, A. Dragan, and R. Shah. Evaluating the robustness of collaborative agents. *arXiv preprint arXiv:2101.05507*, 2021.

[36] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4193–4206, 2017.

[37] A. Lerer and A. Peysakhovich. Learning existing social conventions via observationally augmented self-play. AIES '19, page 107–114, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314268. URL https://doi.org/10.1145/3306618.3314268.

[38] E. Lockhart, N. Burch, N. Bard, S. Borgeaud, T. Eccles, L. Smaira, and R. Smith. Human-agent cooperation in bridge bidding. *arXiv preprint arXiv:2011.14124*, 2020.

[39] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6382–6393, 2017.

[40] R. Lowe, A. Gupta, J. Foerster, D. Kiela, and J. Pineau. Learning to learn to communicate. In *Proceedings of the 1st Adaptive & Multitask Learning Workshop*, 2019.

[41] R. Lowe, A. Gupta, J. N. Foerster, D. Kiela, and J. Pineau. On the interaction between supervision and self-play in emergent communication. *CoRR*, abs/2002.01093, 2020. URL https://arxiv.org/abs/2002.01093.

[42] A. Lupu, H. Hu, and J. Foerster. Trajectory diversity for zero-shot coordination. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 1593–1595, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

[43] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson. Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f816dc0acface7498e10496222e9db10-Paper.pdf.

[44] K. R. McKee, I. Gemp, B. McWilliams, E. A. Duèñez-Guzmán, E. Hughes, and J. Z. Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 869–877, 2020.

[45] K. R. McKee, X. Bai, and S. Fiske. Understanding human impressions of artificial intelligence. Feb 2021. doi: 10.31234/osf.io/5ursp. URL psyarxiv.com/5ursp.

[46] K. R. McKee, J. Z. Leibo, C. Beattie, and R. Everett. Quantifying environment and population diversity in multi-agent reinforcement learning. *arXiv preprint arXiv:2102.08370*, 2021.

[47] F. Melo and A. Sardinha. Ad hoc teamwork by learning teammates' task. *Autonomous Agents and Multi-Agent Systems*, 30, 01 2015. doi: 10.1007/s10458-015-9280-x.

[48] B. Mutlu, A. Terrell, and C.-M. Huang. Coordination mechanisms in human-robot collaboration. In *Proceedings of the Workshop on Collaborative Manipulation, 8th ACM/IEEE International Conference on Human-Robot Interaction*, pages 1–6. Citeseer, 2013.

[49] S. Nikolaidis and J. Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. pages 33–40, 10 2013. ISBN 978-1-4673-3099-2. doi: 10.1109/HRI.2013.6483499.

[50] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving rubik's cube with a robot hand, 2019.

[51] E. Peer, D. M. Rothschild, Z. Evernden, A. Gordon, and E. Damer. MTurk, Prolific or panels? Choosing the right audience for online research. *SSRN*, 2021. doi: 10.2139/ssrn.3765448.

[52] P. M. Pilarski, A. Butcher, M. Johanson, M. M. Botvinick, A. Bolt, and A. S. Parker. Learned human-agent decision-making, communication and joint action in a virtual reality environment. *arXiv preprint arXiv:1905.02691*, 2019.

[53] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991.

[54] S. Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[55] D. Sadigh, S. Sastry, S. Seshia, and A. Dragan. Planning for autonomous cars that leverage effects on human actions. 06 2016. doi: 10.15607/RSS.2016.XII.029.

[56] N. Schurr, J. Marecki, M. Tambe, and P. Scerri. Towards flexible coordination of human-agent teams. *Multiagent and Grid Systems*, 1(1):3–16, 2005.

[57] A. Shih, A. Sawhney, J. Kondic, S. Ermon, and D. Sadigh. On the critical role of conventions in adaptive human-{ai} collaboration. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=8Ln-Bq0mZcy.

[58] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550 (7676):354–359, 2017.

[59] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018. ISSN 0036-8075. doi: 10.1126/science.aar6404. URL https://science.sciencemag.org/content/362/6419/1140.

[60] H. F. Song, A. Abdolmaleki, J. T. Springenberg, A. Clark, H. Soyer, J. W. Rae, S. Noury, A. Ahuja, S. Liu, D. Tirumala, N. Heess, D. Belov, M. Riedmiller, and M. M. Botvinick. V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control. *arXiv preprint arXiv:1909.12238*, 2019.

[61] P. Stone, G. Kaminka, S. Kraus, and J. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.

[62] P. Tylkin, G. Radanovic, and D. C. Parkes. Learning robust helpful behaviors in two-player cooperative atari environments. In *NeurIPS 2020 workshop on Cooperative AI*, 2020. URL https://econcs.seas.harvard.edu/files/econcs/files/tylkin_neurips20.pdf.

[63] P. Tylkin, G. Radanovic, and D. C. Parkes. Learning robust helpful behaviors in two-player cooperative atari environments. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1686–1688, 2021.

[64] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, Oct. 2019. doi: 10.1038/s41586-019-1724-z.

[65] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[66] R. E. Wang, S. A. Wu, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner. Too many cooks: Coordinating multi-agent collaboration through inverse planning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2032–2034, 2020.

[67] F. Wu, S. Zilberstein, and X. Chen. Online planning for ad hoc autonomous agent teams. pages 439–445, 01 2011. doi: 10.5591/978-1-57735-516-8/IJCAI11-081.

[68] S. A. Wu, R. E. Wang, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021. doi: https://doi.org/10.1111/tops.12525. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12525.

[69] S. Zheng, A. Trott, S. Srinivasa, N. Naik, M. Gruesbeck, D. C. Parkes, and R. Socher. The AI economist: Improving equality and productivity with AI-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We claim that our training method (FCP) "scores significantly higher than SP, PP, and BCP when paired with novel agent and human partners" as well as "humans also report a strong subjective preference to partnering with FCP agents over all baselines." These statements are supported by our statistically significant results in Sections 4 and 5.
   (b) Did you describe the limitations of your work? [Yes] See 'Limitations and future work' in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See 'Societal impact' in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The training code is proprietary. The environment code will be made publicly available over the summer independent of this paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See all corresponding figure captions for details.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix D.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See Appendix D.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix D.