Mixed-Initiative Human-Robot Teaming under Suboptimality

Manisha Natarajan[1], **Chunyue Xue**[1], **Matthew Gombolay**[1]

[1] School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
manisha.natarajan@cc.gatech.edu, chunyuexue@gatech.edu, matthew.gombolay@cc.gatech.edu

Abstract

For effective human-agent teaming, robots and other artificial intelligence agents must adapt quickly to their human partner's strengths and preferences. Most prior work typically assume that either the agent or the human acts near-optimally. In real-world collaboration, however, both agents and humans can be suboptimal, due to a lack of information or biases in their decision-making. In this work, we seek to maximize human-robot team performance under suboptimality, where both the human and robot agents are suboptimal to varying extents due to partial knowledge of the environment. We adopt an online Bayesian approach that allows the robot to infer people's preferences and willingness to comply with the robot's assistance in a sequential decision-making game. Our results show that 1) user preferences and team performance can vary with different robot intervention styles, and 2) our proposed Bayesian approach can effectively maximize team performance in a pilot study.

Introduction

Robots are increasingly being deployed to assist humans in highly-demanding, safety-critical scenarios such as emergency response for disasters and terrorist attacks (Nagatani et al. 2013; Casper and Murphy 2003; DeDonato et al. 2015). Robots can assist by taking over tasks to alleviate human workload or acquiring information that complements the human's knowledge. However, both humans and robots have their own limitations while working in such scenarios. For instance, robots may make errors due to high uncertainty in unstructured environments, and humans may make errors from stress and fatigue. In this work, we look at **mixed-initiative autonomy** to maximize human-robot team performance when both agents act suboptimally.

In human-robot teams, mixed-initiative interaction refers to a collaborative strategy where teammates opportunistically seize/relinquish initiative from/to one another during a mission, where initiative can range from low-level motion control to high-level goal specification (Jiang and Arkin 2015). We study mixed-initiative interactions in a teaming task where both the human and the robot only have partial knowledge of the environment and thus act suboptimally. The human and the robot have asymmetric capabilities and must collaborate (seize/relinquish control) to reach a goal location, with the human teleoperating the robot, similar to urban search-and-rescue (USAR) missions (Isaacs et al. 2022). The robot can choose to intervene or override the human's actions, and the human can decide to oppose or comply with the robot. Our goal is to learn a domain-agnostic, online robot policy that determines when and how to intervene to maximize team performance by modeling human latent states. We propose a Bayesian approach that is adaptable to diverse users in mixed-initiative settings.

Achieving ad-hoc, zero-shot coordination with novel human partners has been a longstanding challenge in AI (Klien et al. 2004; Paleja et al. 2021). Recent works aim to achieve human-AI collaboration either from human-human demonstrations (Carroll et al. 2019; Hong, Dragan, and Levine 2023) or via self-play without any human data (Strouse et al. 2021; Zhao et al. 2023). However, these approaches look at domains where both humans and agents have symmetric capabilities. In contrast, our work delves into human-agent teaming with *asymmetric capabilities*, emphasizing the need to actively seize or relinquish control to achieve team objectives. Further, we seek to optimize team performance when all teammates are *suboptimal*, which is seldom explored in human-robot teams (Lee et al. 2020).

We model the human-robot team as a Partially Observable Markov Decision Process (POMDP) (akin to prior works in HRI (Chen et al. 2018; Lee et al. 2020)). The main idea underlying our approach is to learn a robot policy that is conditioned on the uncertainty in the robot's estimation of human behavior. Initially, the robot has high uncertainty about user preferences and willingness to comply. Through Bayesian Learning during interactions, the robot's estimation becomes more refined, reducing its uncertainty. To ensure the feasibility of our approach to run online with novel users, we employ a Monte-Carlo search (scalable to large state spaces) with an approximated belief space and use conjugate priors to perform belief updates efficiently.

The key contributions of the work are two-fold. First, we conduct a user study to show that user preferences and team performance can vary with different robot intervention styles. Next, we develop an adaptive robot policy to intervene users and demonstrate its effectiveness in improving team performance with novel users via a pilot study.

Presented at the AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (AI-HRI) 2023

Preliminaries

We model the human-robot team as a Bayes-Adaptive Partially Observable Markov Decision Process (BA-POMDP) (Ross, Chaib-draa, and Pineau 2007), which enables us to learn a robot policy that dynamically estimates POMDP model parameters during interactions, and is conditioned on the model estimation uncertainty.

Partially Observable Markov Decision Process

A Partially Observable Markov Decision Process (POMDP) is defined as a tuple $\mathcal{M} = (S, A, O, \mathcal{T}, \mathcal{E}, d_0, R, \gamma)$ where S is a set of states $s \in S$, A is a set of actions $a \in A$, O is a set of observations $o \in O$, $\mathcal{T}(s_{t+1}|s_t, a_t)$ is the state transition probabilities, $\mathcal{E}(o_t|s_t)$ is the emission function, d_0 is the initial state distribution, $R(s, a)$ describes the reward of taking action a in state s, and discount factor $\gamma \in (0, 1]$.

The agent's objective in a POMDP is to learn a policy π that maximizes the expected cumulative discounted reward (return). Given the agent's inability to access the true state, it relies on the history of actions and observations, h to learn the policy. Based on h, the agent maintains a probability distribution or belief $b \in \mathcal{B}$ over states and updates the belief with subsequent interactions. Belief updates can be achieved via the Bayes rule (infeasible for large state spaces) or with an unweighted particle filter (approximate update). Hence, the POMDP policy π is a mapping from $\mathcal{B} \rightarrow A$.

Partially Observable Monte-Carlo Planning

Partially Observable Monte-Carlo Planning (POMCP) is an online solver that extends the Monte-Carlo Tree Search (MCTS) to POMDPs (Silver and Veness 2010). POMCP uses a UCT search to select actions and an unweighted particle filter for belief updates. POMCP operates on a search tree of histories h instead of states, where each node in the tree stores statistics – visitation count $N(h)$, value/mean return $V(h)$, and belief $b(h)$. The algorithm performs online planning through multiple simulations, incrementally building the search tree. Each simulation starts by sampling an initial state $s \sim b(h)$. The agent's action selection involves a *tree search policy* within the search tree, optimizing the UCT objective, and a *rollout policy* (often random), outside the search tree. The return of each simulation is used to update the statistics for all visited nodes. Each simulation adds a new node to the tree, corresponding to the first new history encountered in the rollout. POMCP terminates based on preset criteria (e.g., maximum number of simulations).

Bayes Adaptive POMDP

Prior works generally assume that the POMDP is fully specified (i.e., the model parameters $\mathcal{T}, \mathcal{E}$ are known) (Lauri, Hsu, and Pajarinen 2022) which is unrealistic in human-robot collaboration as we neither have access to the true latent states (e.g., trust and preferences) of the human partner nor how these states change during the interaction. Further, estimating $\mathcal{T}, \mathcal{E}$ using maximum likelihood methods do not capture differences across individuals (Chen et al. 2018).

To address these challenges, we adopt the Bayes Adaptive POMDP (BA-POMDP) framework (Ross, Chaib-draa, and Pineau 2007) — a Bayesian Reinforcement Learning approach for solving POMDPs. The BA-POMDP employs Dirichlet count vectors χ to represent uncertainty over parameters $(\mathcal{T}, \mathcal{E})$. Since the POMDP states are hidden, χ cannot be explicitly computed and is also included in the state.

Solving BA-POMDPs is difficult as they are infinite-state POMDPs. Ross et al. introduced an online planner that reduces the BA-POMDPs to finite models (Ross, Chaib-draa, and Pineau 2007). Later Katt et al. proposed the BA-POMCP (an extension of POMCP), the current state-of-the-art, online algorithm for solving BA-POMDPs (Katt, Oliehoek, and Amato 2017). In this work, we use a variant of the BA-POMCP algorithm to optimize performance in suboptimal human-robot teams.

Method

In this section, we first define the human-robot team model for mixed-initiative interactions and then describe how we learn an adaptive robot policy for our current setting.

Human-Robot Team Model

State Space In our human-robot team model, the state space combines the world state and user latent states $s = (x, z)$. The world state $x \in \mathcal{X}$, refers to the task status that the human-robot team is working on, and the latent states $z \in \mathcal{Z}$ can refer to the user's trust or tendency to comply with the robot, and their task execution preferences. The current world state is observable to both humans and robots. We focus on suboptimal human-robot teaming, assuming that the suboptimality arises from task-related errors or incomplete knowledge, i.e., both agents may make errors or cannot observe the full world state. Thus, the world state as observed by the robot may not always align with what the human observes $(x_t^R \neq x_t^H, \forall t)$. The user's latent states are not accessible to the robot, and the robot has to infer these states by observing the user's actions.

Action Space As we are planning from the robot's perspective, the action space encompasses the actions $a^R \in A^R$ that the robot can take in the environment. In our mixed-initiative collaborative scenario, we assume that the robot first observes the human action and then selects its action. The robot can choose to either execute, intervene, or override the user's actions. Additionally, the robot may choose to explain whenever it intervenes or overrides the user.

Observation Space The robot's observations are human actions $a^H \in A^H$. At a high level, the user actions can be categorized as compliance or non-compliance with the robot, i.e., the user can either choose to comply with the robot's last choice or oppose/verify the robot's last choice. We assume that the human's action depends on their knowledge of the current world state x_t and the history of interactions h_{t-1} with the robot, i.e., the human follows the policy – $\pi^H(a_t^H|x_t, h_{t-1})$, where $h_{t-1} = \{a_0^H, a_0^R, a_1^H, a_1^R, \cdots, a_{t-1}^H, a_{t-1}^R\}$. Similar to prior work (Chen et al. 2018), we assume that the user's latent state z_t is a compact representation of the interaction history $(z_t \approx h_{t-1})$. Thus, $\pi^H(a_t^H|x_t, h_{t-1}) \approx \pi^H(a_t^H|x_t, z_t)$.

Transition and Emission Models The transition model $\mathcal{T}$ defines the probability $p(s_{t+1}|s_t, a_t^R)$. In our model, $s_t = (x_t, z_t)$. Thus we can rewrite the transition model as:

$$p(s_{t+1}|s_t, a_t^R) = \sum_{a^H} p(s_{t+1}|s_t, a_t^R, a_t^H) \times \pi^H(a_t^H|x_t, z_t)$$

$$= \sum_{a^H} p(x_{t+1}|x_t, a_t^R, a_t^H) \times p(z_{t+1}|z_t, a_t^R, a_t^H) \times \pi^H(a_t^H|x_t, z_t)$$

The second line in the above equation comes from our assumption that given the user's action, world state dynamics are independent of the user's latent state dynamics. In our collaborative scenario, the world state dynamics are deterministic and known (need not be estimated). Hence, we only estimate the latent state dynamics as part of the BA-POMDP.

The emission model $\mathcal{E}$ for the human-robot team refers to the human policy $\pi^H(a_t^H|x_t, z_t)$ which is also unknown to the robot and must be estimated to solve the BA-POMDP.

Reward Function This work focuses on maximizing performance in a teaming task where both humans and robots can take initiative but are suboptimal. To do so, we consider the reward to be a performance metric for the decision-making task. In our model, the reward is positive for achieving the task goal and negative for failing task constraints.

Adaptive Robot Policy for User Interventions

We adopt a modified version of the Bayes-Adaptive POMCP (BA-POMCP) (Katt, Oliehoek, and Amato 2017) and discuss the key changes we make to this algorithm.

Belief Approximation Similar to POMCP, BA-POMCP constructs a lookahead search tree through environment simulations and maintains a belief over latent parameters using an unweighted particle filter to determine the best action at each time step. However, in BA-POMCP, we need to maintain a belief over both the latent states $|S|$ and the model parameters $\mathcal{T}, \mathcal{E}$ ($|S|^2 \times |A| + |S| \times |A| \times |O|$ parameters). Computing the posterior update over such a large space can be expensive. Further, it is difficult for the posterior distribution to converge to the true parameters, especially when we only have access to limited interactions.

Hence, we leverage the independence assumption between the world state and the latent state transition to approximate the belief in each node in the search tree. Since we only need the human action to determine the next world state, we choose to maintain the belief only over the user action space, instead of all latent states and model parameters. We compute the posterior update for the belief b_{t+1} from the prior belief and interaction history h_t at each node.

Simulating Human Policy In BA-POMCP, we need to simulate human actions during the rollout for constructing the search tree. As the robot lacks direct knowledge of the human policy, we maintain a belief over the space of possible human policies. We model the true human policy as a Bernoulli distribution, with an unknown parameter μ that signifies the likelihood of user compliance with the robot for a given interaction history h. We approximate the belief over the human policy using particles, where each particle acts as a potential candidate for the true human policy. We model each particle as a beta distribution – the conjugate prior for Bernoulli distributions so that the posterior updates can be computed efficiently. To simulate the human action during rollout, we sample a particle from the belief at the current node and use it to estimate the probability of user compliance with the robot. We update the particle (beta distribution) based on interaction experiences, as we continue the simulation until termination.

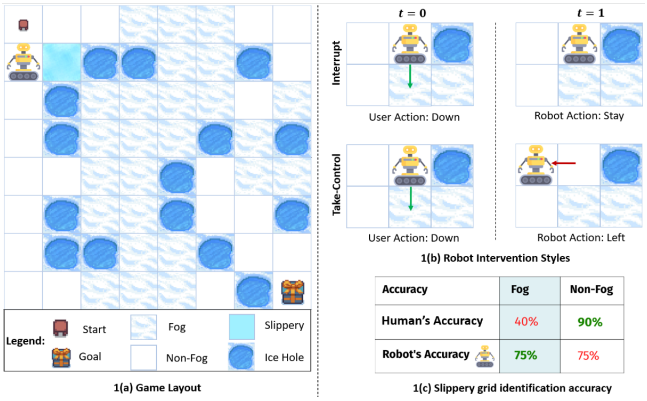


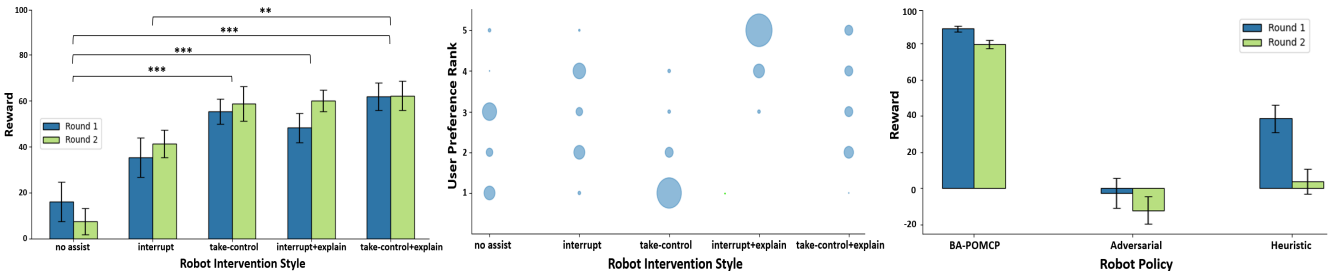
Figure 1: Frozen Lake Domain used in this study. Figure 1(a) shows the overall game layout. Figure 1(b) depicts robot intervention styles, and Figure 1(c) shows the human and robot accuracies for identifying slippery grids.

Human Subjects Experiments

Domain

We modified the Frozen Lake domain from OpenAI Gym (Brockman et al. 2016) for mixed-initiative human-robot teaming. The users must collaborate with a robot to navigate an 8×8 frozen lake grid from start to goal in the fewest steps, while avoiding holes and slippery regions. We modified the original domain to only have certain grids as slippery instead of a uniform slipping probability throughout. Stepping on a slippery region leads to falling into a hole. Both the human and the robot can only observe if adjacent grids are slippery.

To enforce suboptimality, we added errors in the human and robot observations of slippery grids. These errors include – **False Positives** (seeing a safe grid as slippery), and **False Negatives** (seeing a slippery region as safe). Moreover, fog covers some areas of the map, reducing human visibility. The human and robot accuracies for identifying slippery regions are shown in Figure 1. During the game, the human teleoperates the robot across the lake, but the robot may intervene or take control if it finds that the user chose a longer or unsafe path (e.g., slippery regions or holes) to the goal. Additionally, the user is equipped with a high-quality (100% accurate) sensor for detecting slippery regions in adjacent grids, but using it incurs a point cost. The overall performance or game reward is computed as: $90 - \text{steps taken} - 10 \times \#\text{ falls into hole} - 2 \times \#\text{ detections} + 30 \times \mathbb{I}[\text{goal reached}]$. Our environment is inspired by USAR missions, where humans teleoperate robots but humans and robots can have complementary skills and varying domain knowledge.



(a) Performance v/s Intervention (Study 1) (b) Preference v/s Intervention (Study 1) (c) Performance v/s Robot Policy (Study 2)

Figure 2: Results from our Human-Subjects Experiments: Higher reward and higher preference rank is desirable. The error bars imply standard error in Figures 2a and 2c (Statistical significance: ** implies $p < 0.01$ and *** implies $p < 0.001$).

Experiment Design

We propose a two-phase user study design to 1) examine how users respond to different robot intervention styles with and without explanations in a controlled setting (**Data Collection Study**) and 2) evaluate human-robot team performance with our proposed approach (**Evaluation Study**).

Experiment Conditions For the data collection study, we employ a 1×5 within-subjects experiment to examine user responses to various robot interventions. These interventions include – *no assist*: the robot does not intervene (baseline), *interrupt*: the robot stops the user from executing an action, *take-control*: the robot overrides the user's action with its own action, *interrupt+explain*: the robot interrupts and explains, *take-control+explain*: the robot takes over control and explains. To ensure consistency across intervention strategies, the robot employs the same handcrafted heuristic.

In the evaluation study, we employ a 1×3 within-subjects experiment to compare human-robot team performance under different robot policies. The examined policies are our proposed approach, *BA-POMCP*, the *heuristic* policy (used in the data collection study), and an *adversarial* policy (BA-POMCP optimized for inverse reward). We include the adversarial policy as an additional adaptive baseline. To perform a balanced comparison, we ensure that the run times of all policies are identical in the evaluation study.

Metrics For both studies, we assess user preferences and performance using subjective and objective measures respectively. Subjective measures include trust (Muir and Muir 1989), likeability (Bartneck et al. 2009), and willingness to comply (Raemdonck and Strijbos 2013) (adapted from human-human interactions for HRI) measured via 5-point Likert scales. Questionnaires were administered after each round. Demographic data, education, prior robotics experience, and personality are collected via a pre-study questionnaire. At the end of the study, users ranked their preferences for the different robot agents. Objective performance was assessed based on the total reward obtained in each round.

Participants and Procedure We recruited 30 subjects (Age: 25.56 ± 3.38, Female: 33%) for the data collection study and six subjects (preliminary analysis) for the evaluation study from a local university campus after IRB approval. The procedure was the same for both studies. Written consent was obtained before the experiment, and participants received written game instructions along with a demonstration from the experimenter. The subjects had three practice rounds to familiarize themselves and then participated in ten rounds for the data collection study and six rounds for the evaluation study (two rounds per condition). The experiment order was randomized. The subjects were informed to optimize their path to the goal and completed pre- and post-study questionnaires in each round.

Results

Data Collection Study We find that on average the human-robot team performed the best with the *take-control+explain* agent and the least with the *no-assist* condition, proving the necessity for robot interventions (Figure 2a). We used repeated measures ANOVA to obtain statistical results. We did not find statistical significance between intervention styles with and without explanations for performance, but most users preferred to work with agents that provided explanations (Figure 2b).

Evaluation Study We conducted a pilot study ($n = 6$), where we limited the usage of the detection sensor (≤ 5) so as to force the users to rely more on the agent. We found that our proposed adaptive policy outperforms our baselines in terms of human-robot team performance (Figure 2c). We do not report statistical significance as our user population is small. In future work, we will continue to evaluate our algorithm with a larger subject population.

Conclusion

In this work, we propose an online Bayesian approach to optimize performance in human-robot teams when both agents are suboptimal. Our focus is on learning a robot policy for effective user intervention. We find that robot interventions can improve performance while recognizing diverse user preferences. Next, we develop an adaptive robot policy using BA-POMCP and show its effectiveness in improving team performance via a pilot study. We address the computational challenges in BA-POMCP by using a Monte-Carlo search with belief approximation and using conjugate priors to perform belief updates efficiently. In future work, we plan to continue evaluating our algorithm with a larger population and extend it to real-world human-robot collaboration tasks.

References

Bartneck, C.; Kulić, D.; Croft, E.; and Zoghbi, S. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1: 71–81.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI gym. *arXiv preprint arXiv:1606.01540*.

Carroll, M.; Shah, R.; Ho, M. K.; Griffiths, T.; Seshia, S.; Abbeel, P.; and Dragan, A. 2019. On the utility of learning about humans for human-AI coordination. *Advances in neural information processing systems*, 32.

Casper, J.; and Murphy, R. R. 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(3): 367–385.

Chen, M.; Nikolaidis, S.; Soh, H.; Hsu, D.; and Srinivasa, S. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 307–315.

DeDonato, M.; Dimitrov, V.; Du, R.; Giovacchini, R.; Knoedler, K.; Long, X.; Polido, F.; Gennert, M. A.; Padır, T.; Feng, S.; et al. 2015. Human-in-the-loop control of a humanoid robot for disaster response: a report from the DARPA Robotics Challenge Trials. *Journal of Field Robotics*, 32(2): 275–292.

Hong, J.; Dragan, A.; and Levine, S. 2023. Learning to influence human behavior with offline reinforcement learning. *arXiv preprint arXiv:2303.02265*.

Isaacs, J.; Knoedler, K.; Herdering, A.; Beylik, M.; and Quintero, H. 2022. Teleoperation for Urban Search and Rescue Applications. *Field Robotics*, 2: 1177–1190.

Jiang, S.; and Arkin, R. C. 2015. Mixed-initiative human-robot interaction: definition, taxonomy, and survey. In *2015 IEEE International conference on systems, man, and cybernetics*, 954–961. IEEE.

Katt, S.; Oliehoek, F. A.; and Amato, C. 2017. Learning in POMDPs with Monte Carlo tree search. In *International Conference on Machine Learning*, 1819–1827. PMLR.

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

Lauri, M.; Hsu, D.; and Pajarinen, J. 2022. Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, 39(1): 21–40.

Lee, J.; Fong, J.; Kok, B. C.; and Soh, H. 2020. Getting to know one another: Calibrating intent, capabilities and trust for human-robot collaboration. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6296–6303. IEEE.

Muir, B.; and Muir, B. 1989. *Operators' Trust in and Use of Automatic Controllers in a Supervisory Process Control Task*. Canadian theses on microfiche. University of Toronto. ISBN 9780315510142.

Nagatani, K.; Kiribayashi, S.; Okada, Y.; Otake, K.; Yoshida, K.; Tadokoro, S.; Nishimura, T.; Yoshida, T.; Koyanagi, E.; Fukushima, M.; et al. 2013. Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots. *Journal of Field Robotics*, 30(1): 44–63.

Paleja, R.; Ghuy, M.; Ranawaka Arachchige, N.; Jensen, R.; and Gombolay, M. 2021. The utility of explainable AI in ad hoc human-machine teaming. *Advances in neural information processing systems*, 34: 610–623.

Raemdonck, I.; and Strijbos, J.-W. 2013. Feedback perceptions and attribution by secretarial employees: Effects of feedback-content and sender characteristics. *European Journal of Training and Development*, 37(1): 24–48.

Ross, S.; Chaib-draa, B.; and Pineau, J. 2007. Bayes-adaptive POMDPs. *Advances in neural information processing systems*, 20.

Silver, D.; and Veness, J. 2010. Monte-Carlo planning in large POMDPs. *Advances in neural information processing systems*, 23.

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

Zhao, R.; Song, J.; Yuan, Y.; Hu, H.; Gao, Y.; Wu, Y.; Sun, Z.; and Yang, W. 2023. Maximum entropy population-based training for zero-shot human-ai coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6145–6153.