# Mixed-Initiative Human-Robot Teaming under Suboptimality with Online Bayesian Adaptation

Manisha Natarajan*
Georgia Institute of Technology
Atlanta, GA, USA
manisha.natarajan@cc.gatech.edu

Chunyue Xue*
Georgia Institute of Technology
Atlanta, GA, USA
chunyuexue@gatech.edu

Sanne van Waveren
Georgia Institute of Technology
Atlanta, GA, USA
sanne@gatech.edu

Karen Feigh
Georgia Institute of Technology
Atlanta, GA, USA
karen.feigh@gatech.edu

Matthew Gombolay
Georgia Institute of Technology
Atlanta, GA, USA
matthew.gombolay@cc.gatech.edu

## ABSTRACT

For effective human-agent teaming, robots and other artificial intelligence (AI) agents must infer their human partner's abilities and behavioral response patterns and adapt accordingly. Most prior works make the unrealistic assumption that one or more teammates can act near-optimally. In real-world collaboration, humans and autonomous agents can be suboptimal, especially when each only has partial domain knowledge. In this work, we develop computational modeling and optimization techniques for enhancing the performance of human-agent teams, where both the human and the robotic agent have asymmetric capabilities and act suboptimally due to incomplete environmental knowledge. We adopt an online Bayesian approach that enables a robot to infer people's willingness to comply with its assistance in a sequential decision-making game. Our user studies show that user preferences and team performance vary with robot intervention styles, and our approach for mixed-initiative collaboration enhances objective team performance ($p < .001$) and subjective measures, such as user's trust ($p < .001$) and perceived likeability of the robot ($p < .001$).

## KEYWORDS

Human-Agent Teams, Mixed-Initiative, Suboptimality, POMDP

**ACM Reference Format:**
Manisha Natarajan, Chunyue Xue, Sanne van Waveren, Karen Feigh, and Matthew Gombolay. 2024. Mixed-Initiative Human-Robot Teaming under Suboptimality with Online Bayesian Adaptation. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

## 1 INTRODUCTION

Human-agent teaming has the potential to leverage the unique capabilities of humans and artificial intelligence (AI) agents to enhance team performance. However, both humans and agents can be suboptimal, especially under uncertainty [19, 23]. Imagine a human collaborating with a robot in an urban search-and-rescue (USAR) mission with reduced visibility due to fog or smoke. The human can take control when the robot is error-prone (e.g., in unstructured environments). Likewise, the robot can intervene when human vision is limited. Optimizing this collaboration requires robots to develop a Theory of Mind [35], i.e., the ability to infer the human teammates' mental states and anticipate their actions to determine when intervention is beneficial. In this work, we look at *mixed-initiative interactions*, where the robot models human behavior to decide when to intervene to maximize team performance.

In human-robot teams, mixed-initiative interaction refers to a collaborative strategy in which teammates opportunistically seize and relinquish initiative from and to each other during a mission, where initiative can range from low-level motion control to high-level goal specification [18]. We study such interactions in a teaming task where humans and robots act suboptimally due to partial environmental knowledge. Specifically, the human teleoperates the robot, similar to USAR missions [16], and must collaborate with the robot (seize or relinquish control) to reach a goal location. The human and the robot have asymmetric capabilities and non-identical, partial knowledge of the environment. During the task, when the human selects an action, the robot can comply, interrupt, or take over with an alternative action. The human can then decide to accept or oppose the robot's decision.

Our goal is to learn a domain-agnostic robot policy that can effectively adapt to diverse users to maximize team performance without prior human interaction data. Achieving such ad-hoc or zero-shot coordination with novel human partners has been a long-standing challenge in AI [22, 33]. Recent works explore zero-shot human-AI collaboration by learning AI agent policies either from human-human demonstrations [5, 14] or via self-play without any human data [45, 47]. 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*, where mixed-initiative teaming is essential. While prior works in mixed-initiative teaming have adopted strategies for switching control between humans and robots by estimating performance [7, 38], or operator engagement [9], our work differs by explicitly modeling user compliance to determine when robots should intervene to maximize performance.

Our contributions are two-fold. First, we propose a novel, online, Bayesian approach for zero-shot human-robot collaboration

*Both authors contributed equally to the paper.

This work is licensed under a Creative Commons Attribution International 4.0 License.

*Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand.* © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

in mixed-initiative settings. We assume no prior knowledge of human capabilities. We model the human-robot team as a Partially Observable Markov Decision Process (POMDP), where the robot maintains a belief over users' compliance tendencies. Initially, the robot has high uncertainty about user compliance, but Bayesian updates during subsequent interactions refine the estimation, reducing the uncertainty. By conditioning the robot's policy on the estimated individual's compliance, our approach is more robust to adapt to a diverse pool of participants than having a single, unified model for all subjects. To address the computational challenges in solving POMDPs and ensure that our approach is feasible to run online, we employ a Monte-Carlo search while anticipating appropriate user behavior with approximate belief updates.

Second, we design a new user study interface for examining mixed-initiative human-robot teaming. We open-source our implementation[1]. Through two human-subjects experiments ($n = 30$ and $n = 28$), we demonstrate that (1) user preferences and team performance can vary when the robot employs different intervention styles, and (2) our proposed approach performs favorably on both objective (team performance) and subjective (users' trust, robot likeability) metrics with novel users.

## 2 RELATED WORK

### 2.1 Modeling Human Behavior

For seamless human-robot collaboration, robots must anticipate human behavior and act accordingly. Prior works have shown that robots modeling human behavior can enhance team performance in various applications, such as autonomous driving [40], assistive robotics [17], and collaborative games [34]. Both model-free and model-based approaches have been employed for modeling human behavior. Model-free approaches (e.g., imitation learning [5] and inverse reinforcement learning [13]) require substantial data and generally employ neural networks to learn human behavior.

In contrast, model-based approaches require far fewer samples but make certain assumptions about human behavior (e.g., humans exhibit bounded rationality [44]). In Human-Robot Interaction (HRI), POMDPs and their variants (e.g., BAMDP, MOMDP, I-POMDP) are often used to account for latent factors such as trust, intent, or capability influencing human decision-making [6, 25, 37, 46]. However, existing works mostly assume known model parameters or use maximum likelihood estimation (MLE) [6, 26, 37, 46], which may not generalize well across individuals and can overfit [3]. Hence, we instead adopt a Bayesian approach to jointly learn the POMDP parameters and the robot policy during human interactions, similar to prior work [25, 29, 32]. Our work differs from prior Bayesian approaches in HRI by maintaining belief about *dynamic* latent parameters, such as trust or compliance [6, 30], which vary during interactions and across individuals. To address the computational complexity of Bayesian approaches, we approximate the belief space and use conjugate priors for efficient belief updates.

### 2.2 Human-Agent Teaming

Recently, there has been a surge in interest in designing AI agents that are capable of collaborating with humans, especially in ad-hoc settings [1, 14, 15]. Ad-hoc or zero-shot human-agent teaming requires agents to be adept at collaborating with diverse users in novel contexts without prior interactions. Achieving ad-hoc, zero-shot coordination with novel human partners has been a longstanding challenge in AI and will be crucial for the ubiquitous deployment of robots and AI agents [22, 31, 33]. Recent works aim to achieve ad-hoc human-AI teaming either from human-human demonstrations using Behavior Cloning [5] and offline RL [14] or via self-play without any human data [45]. Others have also explored population-based training to learn robot policies that are generalizable across diverse users [27, 47]. However, these approaches focus on domains where both humans and agents have symmetric capabilities and work concurrently. In contrast, our work examines mixed-initiative teaming, where humans and agents must assume or yield control to achieve the task objective. Mixed-initiative approaches have been previously employed in human-agent teams to share control over high-level goals or task specification [11, 12], as well as low-level motion control of agents [7, 41]. The initiative is considered **mixed** only when each team member is authorized to intervene or seize control [18]. In our study, we investigate mixed-initiative motion control in scenarios when humans and agents possess *asymmetric capabilities*. Hence, we cannot apply approaches from prior work for learning robot policies from human-human demonstrations. Instead, we develop an online algorithm for learning the robot's intervention policy to improve team performance.

## 3 PRELIMINARIES

We model the human-robot team as a Bayes-Adaptive POMDP (BA-POMDP) [39], allowing the robot to dynamically learn and adjust its policy based on estimations of human model parameters while accounting for estimation uncertainty.

A 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, $R(s_t, a_t)$ is the reward, and $\gamma \in (0, 1]$ is the discount factor. The agent's goal is to learn a policy, $\pi : \mathcal{B} \rightarrow A$, that maximizes the expected cumulative discounted reward (return), where $b \in \mathcal{B}$ is a belief state inferred by a history of previous observations and actions, $h$. Belief updates can be achieved via the Bayes rule (infeasible for large state spaces) or with an unweighted particle filter (approximate update).

Most prior works in POMDPs assume a fully specified environment (i.e., the model parameters $\mathcal{T}$, $\mathcal{E}$ are known) [24], which is unrealistic in HRI as we neither have access to the person's true latent states (e.g., trust, intent) nor how they change during the interaction. We adopt the BA-POMDP framework — a Bayesian Reinforcement Learning approach for solving POMDPs [39]. The BA-POMDP employs Dirichlet vectors, $\chi$, to represent uncertainty over the model parameters $(\mathcal{T}, \mathcal{E})$. As the POMDP states are hidden, $\chi$ cannot be computed and is considered as part of the hidden state.

### 3.1 Solving POMDPs

Partially Observable Monte-Carlo Planning (POMCP) is an online solver for POMDPs that uses the Monte-Carlo Tree Search (MCTS) [43]. POMCP uses the UCT (Upper Confidence Bound (UCB) for

[1] https://github.com/CORE-Robotics-Lab/Bayes-POMCP

Trees) to select actions and an unweighted particle filter for belief updates. In POMCP, the UCT algorithm is extended to partially observable domains using a search tree of histories $h$ instead of states, where each node in the tree stores statistics – visitation count, $N(h)$, value or mean return, $V(h)$, and belief, $b(h)$, approximated by particles. The algorithm performs online planning through multiple simulations, incrementally building the search tree. The return of each simulation is used to update the statistics for all visited nodes. POMCP terminates based on preset criteria (e.g., max simulations).

We model the human-robot team as a BA-POMDP. Solving BA-POMDP models is difficult as they are infinite-state POMDPs. The current state-of-the-art online solver for BA-POMDPs is the BA-POMCP (i.e., POMCP for BA-POMDPs) [20]. In this work, we propose **Bayes-POMCP**, which extends the BA-POMCP algorithm for human-robot teams, by incorporating belief approximations and simulating user actions for improved search efficiency.

## 4 METHOD

In this section, we first define the human-robot team model (BA-POMDP) for mixed-initiative interactions and then describe our approach, **Bayes-POMCP**, to learn an adaptive robot policy for mixed-initiative human-robot teams.

### 4.1 Human-Robot Team Model

*4.1.1 State Space.* In our human-robot team model, the state space combines the world state and user latent state, $s = (x, z)$. The world state, $x \in \mathcal{X}$, refers to the task at hand, and the latent state, $z \in \mathcal{Z}$, refers to the user's trust or tendency to comply with the robot. The robot does not have access to the user's latent state and must deduce it from user actions. In this work, we assume that the world state dynamics are independent of the user's latent state, given the user's action. We focus on suboptimal human-robot teaming, assuming that the suboptimality arises from incomplete knowledge, i.e., both agents may make errors as they cannot observe the full world state. Thus, the world state as observed by the robot may not always align with what the human observes ($x^R_t \neq x^H_t, \forall t$).

*4.1.2 Action Space.* As we are planning from the robot's perspective, the action space comprises 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[2]. 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.

*4.1.3 Observation Space.* The robot observes the human actions $a^H \in A^H$. 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, $\pi^H(a^H_t|x_t, h_{t-1}, a^R_{t-1})$, where $h_{t-1} = \{a^H_0, a^R_0, a^H_1, a^R_1, \cdots, a^H_{t-1}\}$. Similar to prior work [6], we assume that the user's latent state, $z_t$, is a compact representation of the interaction history ($z_t \approx \{h_{t-1} \cup a^R_{t-1}\}$). Thus, $\pi^H(a^H_t|x_t, h_{t-1}, a^R_{t-1}) \approx \pi^H(a^H_t|x_t, z_t)$.

[2]Our approach is not restricted to this mixed-initiative setting and can be extended to cases where either the robot takes the first action or works concurrently with users.

*4.1.4 Transition and Emission Models.* We define the state transition model, $\mathcal{T}$, from the robot's perspective, i.e., $\mathcal{T} = p(s_{t+1}|s_t, a^R_t)$. However, for mixed-initiative settings, the state transitions occur as a result of both human and robot actions (Equations 1, 2):

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

$$= \sum_{a^H} p(x_{t+1}|x_t, a^R_t, a^H_t) \times p(z_{t+1}|z_t, a^R_t, a^H_t) \times \pi^H(a^H_t|x_t, z_t) \quad (2)$$

Equation 2 comes from our assumption that given the human and robot actions, the world state dynamics are independent of the human latent state dynamics. In this work, we only estimate the latent state dynamics as part of the BA-POMDP, as we assume that the world state dynamics are deterministic and known.

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

*4.1.5 Reward Function.* The reward function $\mathcal{R}(x, a^H, a^R)$ is positive for team actions that aid in achieving the task objective and negative for team actions that hinder task success. We assume that both the user and the robot know the reward function.

### 4.2 Adaptive Robot Intervention Policy

To maximize human-robot team performance in real-time for mixed-initiative settings, we implement a modified version of the BA-POMCP [20]. Here, we highlight the key changes we make to the BA-POMCP algorithm. Figure 1 provides an overview of our approach, and the complete procedure is described in Algorithm 1.

**Algorithm 1: Bayes-POMCP: Maximizing Performance in Mixed-Initiative Human-Robot Teams**

**Input:** Initial world state $x_0$; Interaction history $h_{-1} = []$; initial belief $b_0$; Search Tree $T = \{\}$

$a^R_{-1} \leftarrow$ No-Assist // By default before episode starts
$z_0 \leftarrow \{h_{-1} \cup a^R_{-1}\}$ // Initial human latent state
$a^H_0 \leftarrow \textsc{RealHuman}(\cdot|x_0, z_0)$ // First human action
$h_0 \leftarrow [a^H_0]$
$T(h_0) \leftarrow \textsc{ConstructNode}(T, h_0)$ // Construct root node
**for** $t = 0, 1, 2, \ldots$ *max_steps* **do**
  $a^R_t \leftarrow \textsc{Search}(h_t)$ // Root node ▷ Search (Supp. Alg. 2)
  **if** $(h_t a^R_t) \notin T$ **then**
    $\textsc{ConstructNode}(T, h_t a^R_t)$
  $x_{t+1} \leftarrow p(\cdot|x_t, a^R_t, a^H_t)$ // Update World State
  $z_{t+1} \leftarrow \{h_t \cup a^R_t\}$ // Update true latent state ⇏ Robot
  $a^H_{t+1} \leftarrow \textsc{RealHuman}(\cdot|x_{t+1}, z_{t+1})$ // Next user action
  $h_{t+1} \leftarrow h_t \cup \{a^R_t, a^H_{t+1}\}$
  **if** $h_{t+1} \notin T$ **then**
    $T(h_{t+1}) \leftarrow \textsc{ConstructNode}(T, h_{t+1})$
  $b(h_{t+1}) \leftarrow \textsc{Belief-Update}(b(h_t), a^R_t, a^H_{t+1})$
  $\textsc{Prune-Tree}(T, h_{t+1})$ // $h_{t+1}$ is the root node

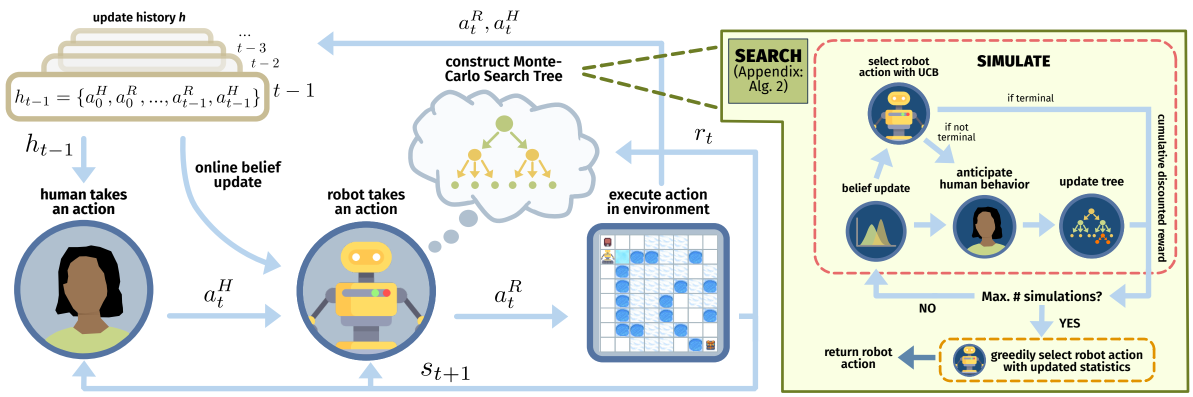



**Figure 1: Graphical overview of the Bayes-POMCP approach for mixed-initiative Human-Robot Teaming: At each timestep $t$, the human first takes an action based on interaction history, $h$, and their current observation of the world state, $x$. The robot then determines when and how to intervene by anticipating human behavior using a Monte-Carlo tree search. The reward is calculated based on both human and robot actions.**

*4.2.1 Belief Approximation.* Similar to POMCP, BA-POMCP also constructs a search tree via environment simulations and maintains a belief over latent parameters using an unweighted particle filter to determine the best action at each time step. However, BA-POMCP requires maintaining a belief over both the latent states $|S|$ and the model parameters $\mathcal{T}, \mathcal{E}$ (i.e., $|S|^2 \times |A| + |S| \times |A| \times |O|$ parameters). Computing the posterior update over such a large space can be expensive, and achieving convergence to true parameters is challenging, especially with limited interactions.

Hence, we leverage the independence assumption between the world state and the latent state transition (Equation 2) to approximate the belief in each node in the search tree, making it feasible to compute the belief updates in real-time. Since only the human action is needed to determine the next world state, we choose to maintain the belief only over user compliance. We compute the posterior update for the belief $b(h_{t+1})$ from the prior belief, $b(h_t)$, based on the interaction history, $h_t$, at each node.

*4.2.2 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 true human policy, we first estimate the human policy parameters and use the same for simulation. We model the true human policy as a Bernoulli distribution with an unknown parameter, $\mu$, that signifies the likelihood of user compliance for a given interaction history, $h$. To estimate $\mu$, we adopt a Bayesian approach. We assume a prior distribution or belief over the space of human policies $b = p(\mu)$. We approximate $b$ using a set of particles, which is updated upon subsequent interactions with the user. Updating beliefs can be computationally expensive, but for the conjugate family of distributions, such updates can be computed efficiently [3]. Thus, we model each particle as a beta distribution – the conjugate prior for Bernoulli distributions.

During the rollout, we simulate human actions by sampling a particle from the current belief, $b$. This sampled particle informs whether the user will comply with the robot's interventions. Additionally, we assume that humans are rational and employ an $\epsilon-$greedy heuristic to select the user's actions in case of noncompliance. The belief, $b$, is updated based on the interaction outcomes during simulation. For further details, refer to the Supplementary.

Alternatively, we can use a random policy to mimic human behavior, but this would require more simulations to cover a range of possible human responses and determine the optimal robot action—resulting in increased computation time. Thus, we opt for estimating user compliance and then simulating the human actions, which we find empirically to be more efficient. To evaluate the contributions of our proposed modifications to the BA-POMCP algorithm [20], we perform an ablation analysis without modeling humans. We refer to this approach as POMCP in our analysis (Section 6.2).

# 5 EVALUATION

## 5.1 Domain

We modified the Frozen Lake environment from OpenAI Gym [4] for evaluating mixed-initiative human-robot teaming. In this domain, the users must collaborate with the robot to navigate an $8 \times 8$ frozen lake grid from start to goal in the fewest steps possible while avoiding holes and slippery regions. We modified the original domain to only have certain grids as slippery instead of a constant slip probability throughout the map. Stepping on a slippery region will cause the agent to fall into a hole. Both the human and the robot can only observe whether the adjacent four grids are slippery. Each time the agent falls into a hole, the team incurs a penalty $\alpha$ and must begin again from the start location.

To enforce suboptimality, we introduce errors in the human and robot observations of slippery grids. These errors include – **False Positives** (observing a safe grid as slippery), and **False Negatives** (observing a slippery region as safe). Moreover, certain parts of the map are covered by fog which reduces human visibility. The human and robot accuracies for identifying slippery regions are shown in Figure 2. 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

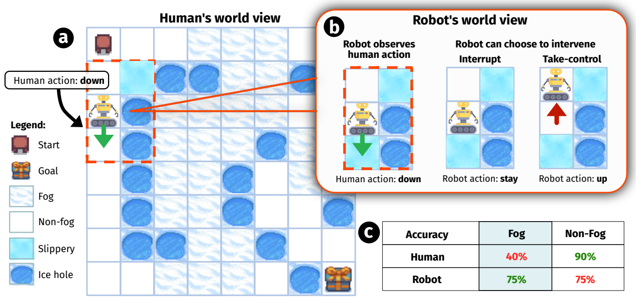



**Figure 2: Frozen Lake Domain used in this study. Figure 2(a) shows the overall game layout. Figure 2(b) depicts robot intervention styles: interrupt, take-control, and Figure 2(c) shows the human and robot accuracies in identifying slippery grids.**

a high-quality (100% accurate) sensor for detecting slippery regions in adjacent grids, but each use of the sensor incurs a point cost $\rho$. The overall team performance or game reward for each round is calculated as a combination of step penalty (shorter path $\longrightarrow$ higher reward), penalty for falling into holes $\alpha$, detection penalty $\rho$, and a bonus $\kappa$ for reaching the goal as shown in Equation 3.

$$\text{Reward} = \text{Max steps} - \#\text{ steps taken} - \alpha \times \#\text{ falls into hole} \\ - \rho \times \#\text{ detections} + \kappa \times \mathbb{1}[\text{goal reached} == \text{True}] \quad (3)$$

We empirically set max steps = 80, $\alpha = 10$, $\rho = 2$ and $\kappa = 30$ for our human-subject experiments. Our environment is inspired by USAR missions, where humans teleoperate robots, but both humans and robots can have complementary skills and varying domain knowledge. Further details of the user study domain can be found in the Supplementary.

## 5.2 Human-Subjects Experiments

We conducted two user studies to 1) examine how users respond to different robot intervention styles with and without explanations but with a static policy (**Data Collection Study**) and 2) evaluate human-robot team performance with the proposed adaptive Bayes-POMCP approach (**Evaluation Study**).

*5.2.1 Data Collection Study.* We employ a $1 \times 5$ within-subjects experiment design to examine user responses to various robot interventions in mixed-initiative teaming (Figure 2b). 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 that determines when to intervene. The heuristic intervention policy is a short-horizon planner that only intervenes if the user's current action is anticipated to lead to a slippery region (based on the robot's knowledge), a hole, or a longer path (≥k steps) and will cede control to the user if the user persistently chooses the action the robot is intervening. The heuristic employs a static intervention style. The algorithm for the heuristic policy can be found in the Supplementary.

*5.2.2 Evaluation Study.* We employ a $1 \times 3$ within-subjects experiment to compare human-robot team performance under different robot policies. The examined policies are our proposed approach – Bayes-POMCP, the same heuristic policy as was used in the data collection study, and an adversarial policy (Adv-Bayes-POMCP) optimized for negative game reward (Equation 3). We include the adversarial policy as an adaptive baseline to show that (1) our proposed approach can successfully aid or inhibit the user from reaching the goal, and (2) it is essential for the adaptive policy to reason when to intervene effectively in addition to switching the intervention styles. To perform a balanced comparison, we ensure that the run times of all robot policies are identical. Further, we limit the use of the detection sensor ($\leq 5$) in the evaluation study to force participants to rely on the robot's assistance.

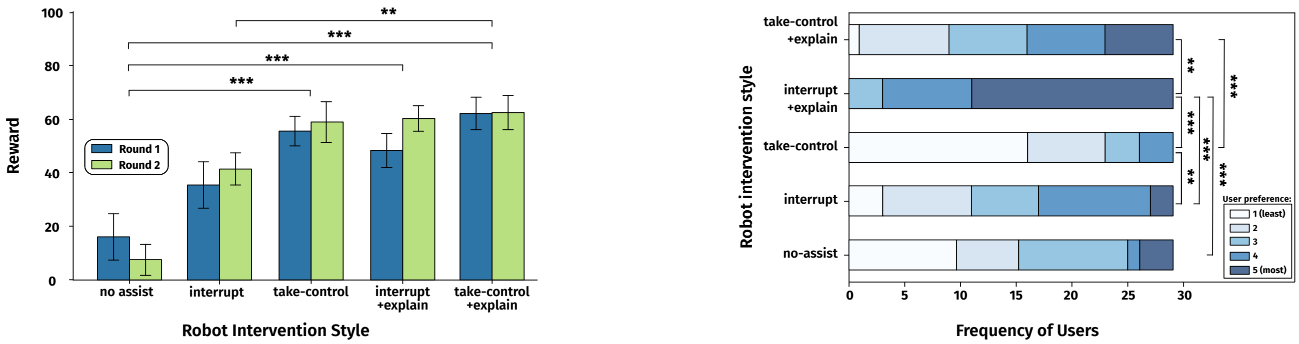
*5.2.3 Metrics.* For both studies, we assess user preferences and performance using subjective and objective measures, respectively. Our subjective measures include trust [28], likeability [2], and willingness to comply [36] (adapted from human-human interactions for HRI) measured via 5-point Likert scales. All questionnaires were administered to the users after each round in both studies. Further, participants reported their demographics, highest completed education, prior experience with robots, and completed a 50-item personality scale [10] as part of the pre-study questionnaire. At the end of the study, users ranked their preferences for the different robot agents. All questionnaires used for the study can be found in the Supplementary. Objective performance was assessed based on the total game reward (Equation 3) in each round.

*5.2.4 Participants and Procedure.* We recruited 30 participants (Age: $25.56 \pm 3.38$, Female: 33%) for the data collection study and 28 new participants (Age: $25.27 \pm 3.28$, Female: 50%) for the evaluation study, all from a local university campus after IRB approval. The procedure was the same for both studies. Written consent from the participants was obtained before the experiment. At the start of the study, participants received written game instructions along with a demonstration from the experimenter. Participants first completed three practice rounds to familiarize themselves with the game and then engaged in ten and six rounds (two rounds per condition) for the data collection study and evaluation study, respectively. The subjects were instructed to complete each round by taking the shortest path to the goal. The experiment order was randomized, and participants completed pre- and post-study questionnaires.

## 5.3 Hypotheses

To investigate how different users interact with various robot intervention styles, we first conducted a data collection study. We hypothesize the following:

**H1A:** *The human-robot team performance can vary with different robot intervention styles.* Although the robot follows the same heuristic across different conditions in Study 1, we hypothesize that the team performance will vary across intervention styles as users may respond differently. For instance, users may be better informed to choose the next action appropriately when the robot intervenes and provides an explanation.




(a) Team Performance vs. Robot Intervention Style

(b) User preferences for working with different robots

**Figure 3: Results from the Data Collection Study with Heuristic Mixed-Initiative Policies. Figure 3a shows that users perform the best with the *take-control* agents and the worst with the no-assist (baseline) agent. Figure 3b depicts users' preference rankings, with the *interrupt+explain* agent being the most favored (rank = 5). Error bars indicate standard error.**

**H1B:** *Users will have different preferences for working with various robot intervention styles.* Humans have varying personality traits and task preferences, which may impact how they perceive and collaborate with teammates. For instance, extroverted individuals are more likely to assume leadership and less likely to renounce control in human-human teams [21]. Likewise, we hypothesize that users will have different preferences when working with robots that interrupt or take control with or without offering explanations.

For the evaluation study, we compare the team performance with the Bayes-POMCP policy against heuristics used in the first study and an adversarial baseline – Adv-Bayes-POMCP. We hypothesize:

**H2A:** *The human-robot team performance will be the highest when the robot employs the Bayes-POMCP policy.* We hypothesize that the Bayes-POMCP policy, which actively anticipates human actions by considering their latent states, is better suited for determining when and how to intervene various users and will thereby maximize team performance. In contrast, the baselines that do not model the human latent states (the heuristic policy) or optimize for negative reward (the adversarial Bayes-POMCP) will perform poorly.

**H2B:** *Users will most prefer to work with our proposed approach, the adaptive Bayes-POMCP policy.* We hypothesize that the Bayes-POMCP policy can effectively intervene users by modeling their latent states and will, therefore, not only improve team performance but also have a positive impact on the users' subjective preference for collaborating with the robot.

# 6 RESULTS AND DISCUSSION

In this section, we first discuss the results of the data collection study. Next, we show results from our simulation experiments used to validate Bayes-POMCP before testing on human participants. We then discuss the results from the evaluation study, comparing our Bayes-POMCP approach and two baselines.

All our statistical analyses were performed using libraries in R, and the significance level, $\alpha$, was set at 0.05. For our analysis, we use parametric tests unless the model fails to meet the required assumptions (e.g. normality, homoscedasticity).

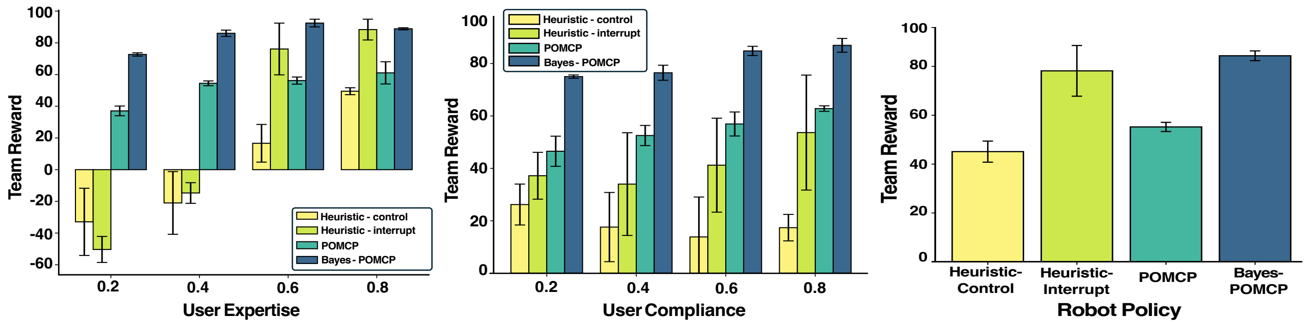
## 6.1 Data Collection Study

For the data collection study, we recruited 30 participants and excluded one participant as an outlier since they failed to complete all ten rounds in the study (failure rate across all subjects: $1.733 \pm 2.365$). Thus we have data from 29 subjects for our analysis.

**H1A:** *Team Performance and Robot Intervention Styles.* We compare the team performance using the game reward (Equation 3) across the five robot intervention styles employed in the first study. The robot either used the same heuristic policy to determine when to intervene or did not intervene at all (no-assist: baseline condition). Each user participated in two rounds for each intervention style, totaling ten rounds; all played on different maps with varying levels of difficulty. To mitigate ordering effects and map-related biases, the experiment conditions and map assignments were randomized. We use Kruskal-Wallis (a non-parametric test), with the dependent variable as the reward and the independent variable as the robot intervention style. We obtain statistical significance for the intervention style ($H(4) = 58.16, p < .001$). We use Dunn's test with Holm-correction for performing post-hoc pairwise comparisons, and the significance values are shown in Figure 3a.

**Takeaway:** We find that the human-robot team performance is impacted by the intervention styles used by the robot, rejecting the null hypothesis (Figure 3a). Firstly, it is worth noting that the team performance significantly improves when the robot intervenes compared to the baseline (no assistance), validating the need for robot interventions in our study domain. Secondly, the team performance is the highest when the robot takes over control. Lastly, adding explanations did not significantly improve performance for the same intervention style (e.g., between interrupt and interrupt+explain).

**H1B:** *Users' Working Preference and Robot Intervention Styles.* At the end of the first user study, participants were asked to rank their preferences for working with various robot intervention styles on a scale from 1 (lowest) to 5 (highest). As user rankings are considered ordinal data, we use Kruskal-Wallis, a non-parametric test, to analyze **H1B**. We find that robot intervention style indeed influences user preferences ($H(4) = 61.67, p < .001$). The majority




**(a) Static Users - vary expertise ($\psi$).** **(b) Static Users - vary compliance ($\theta$).** **(c) Dynamic Users**

**Figure 4: Team performance in simulation experiments with static and dynamic latent user models. Figures 4a and 4b show that Bayes-POMCP can enhance team performance across users of varied expertise and compliance tendencies, respectively. Bayes-POMCP outperforms heuristics and the ablation POMCP model, especially for users with low expertise.**

of the users preferred the interrupt+explain agent the most and the take-control agent the least, as shown in Figure 3b.

**Takeaway:** Our results suggest that, despite explanations not improving performance, most users favor working with robots that offer explanations for their interventions. Interestingly, even though the take-control agent achieved the highest team performance, it was the least preferred choice for the majority of users. These findings highlight the need for an adaptive robot policy that adjusts the intervention style to maximize performance and user satisfaction. If the robot only takes over control, it can improve team performance in the short term but can cause user dissatisfaction and can potentially lead to users abandoning the system in the long run.

## 6.2 Simulation Experiments

Before testing the Bayes-POMCP policy on actual users, we first ensure its adaptability by experimenting with various simulated human models. In these simulation experiments, we compare Bayes-POMCP against two baselines – (1) the standard POMCP algorithm [43] with no human model (POMCP) and (2) the heuristic agents (both take-control and interrupt) on five of the 8×8 maps used in the data collection study. To simulate a diverse set of users, we modulate two latent parameters that determine their behavior – the users' capability or expertise ($\psi$) to solve the task and the users' tendency to comply with the agent ($\theta$). We test with both static users (whose latent parameters – $\psi, \theta$ are fixed) and dynamic users, whose $\theta$ varies continuously based on the interaction history, but $\psi$ remains fixed (i.e., we assume no learning effect as the domain is simple). We provide further details of the simulated human population in the Supplementary. Our results (Figure 4) indicate that Bayes-POMCP outperforms both the heuristics employed in the first study and the ablation baseline without human modeling (POMCP) for static and dynamic user models.

## 6.3 Evaluation Study

Upon verifying our policy with simulated models, we collected data from 28 new participants (who did not take part in the first user study) for the evaluation study. We excluded data from three subjects. Two of the three subjects encountered graphic rendering issues in the study interface. The third subject failed to complete all six rounds (failure rate across all subjects: 3.48 ± 0.77).

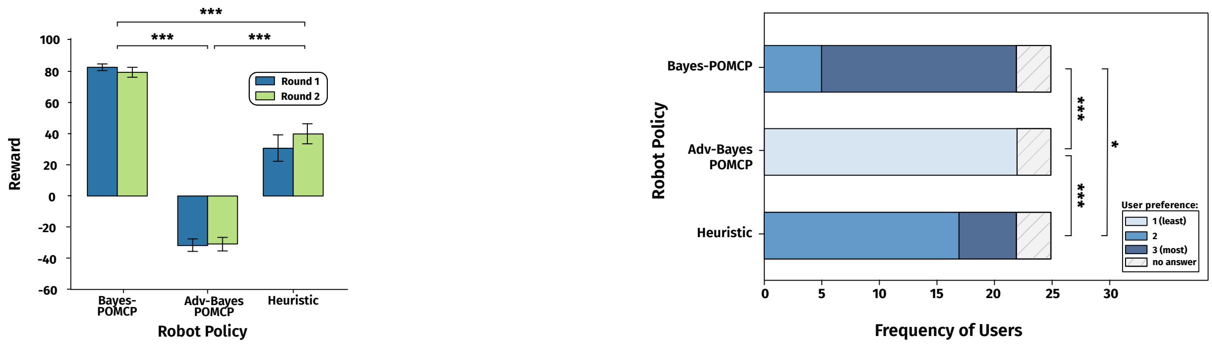
**H2A:** *Team Performance and Robot Policy.* We evaluated the team performance for different robot policies – the heuristic agents (interrupt+explain and take-control+explain) from the first study, our proposed approach, Bayes-POMCP, optimized for the true and negative reward. Each user participated in two rounds per policy, totaling six rounds, played on different maps (a subset from the first study). We used the Kruskal-Wallis test with the reward as the dependent variable and the robot policy as the independent variable. We obtained statistical significance for the robot policy ($H(2) = 109.89, p < .001$). Post-hoc analysis was conducted using Dunn's test with Holm-correction, as shown in Figure 5a.

**Takeaway:** We find that Bayes-POMCP policy significantly outperforms our baselines for team performance (Figure 5a). We also find that the adversarial Bayes-POMCP is effective in preventing the user from reaching the goal, as reflected by the negative reward.

**H2B:** *Users' Working Preference and Robot Policy.* Users ranked their preferences for working with the different robot agents at the end of the second study. We perform the Kruskal-Wallis test, which shows that the robot policy significantly influences user preferences ($H(2) = 45.41, p < .001$). We find that 68% of the users preferred the Bayes-POMCP agent the most, 20% preferred heuristic agents the most, and 88% preferred the Adv-Bayes-POMCP agent the least. 12% (= 3/25) did not answer the preference survey.

We also analyzed subjective metrics with Likert scales for trust, willingness to comply, and robot likeability. We conducted three rANOVA with the subjective metrics as the dependent variables and independent variables as robot policy, number of rounds completed, demographics (age, gender, prior robotics experience), and pre-study questionnaire responses of the user. We find that robot policy was statistically significant across all subjective metrics from the three ANOVAs, with our proposed approach having the highest mean values. We then performed post-hoc analysis using Tukey HSD. For further details of the analysis, see Supplementary.

**Takeaway:** We find that the Bayes-POMCP policy significantly outperforms our baselines across all subjective metrics, and the majority (68%) preferred to work with the Bayes-POMCP agent.




(a) Team Performance vs. Robot Policies

(b) User Working Preferences for Robot Policies

**Figure 5: Results from the Evaluation Study. Figure 5a shows that the team performance is the highest for the *Bayes-POMCP* agent and the lowest for the Adv-Bayes-POMCP (the adversarial baseline). Figure 5b shows that the majority of the users prefer our approach compared to the baselines. Error bars indicate standard error.**

| Hypotheses | I.V. | Levels | n | D.V. | Effect Size | Power |
|---|---|---|---|---|---|---|
| H1A | Robot Intervention Style | 5 | 29 | Reward | 0.201 | 0.3728 |
| H1B | Robot Intervention Style | 5 | 29 | User Ranking | 0.428 | 0.995 |
| H2A | Robot Policy | 3 | 25 | Reward | 0.738 | 0.923 |
| H2B | Robot Policy | 3 | 25 | User Ranking | 0.528 | 0.581 |

**Table 1: Power analysis: D.V. and I.V. refer to the dependent and independent variables, and $n$ is the number of subjects.**

## 6.4 Power and Effect Size Analysis

For all our hypotheses, **H1A, H1B, H2A, H2B**, we used non-parametric tests as the dependent variable was ordinal data or the model did not pass the parametric test assumptions. We report the effect size and statistical power for our analyses in Table 1. Tests for **H1A** and **H2B** are underpowered, warranting additional data for increased confidence in our results. We note that **H2B** might be underpowered due to the challenges in measuring user preferences for adaptive systems [8] and posit that solely relying on post-trial surveys may be insufficient.

# 7 LIMITATIONS AND FUTURE WORK

There are limitations to our experimental findings. Our results were validated in a grid world with discrete actions using university students. Further research is warranted in more complex environments with a larger, diverse user population. We also find that assessing user preferences for adaptive policies (**H2B**) with post-trial surveys may be insufficient, and additional avenues for collecting user preferences during interactions should be explored in future work.

While Bayes-POMCP successfully improves human-robot team performance in a computationally efficient manner, it relies on an environment simulator to estimate the value of human-robot actions in the Monte Carlo search tree, which may not be available for real-world human-robot collaboration tasks. Thus, we aim to learn evaluation functions as shown in prior work [42]. Moreover, our findings indicate that robot explanations only influenced users' subjective perceptions but did not impact team performance. We hypothesize that this may be because the task was relatively simple, and users did not need explanations from the robot to enhance their decision-making. In future work, we seek to assess the utility of explanations in improving team performance by extending our approach to use robots in realistic settings [7], thereby increasing the problem complexity (e.g., complex human models, larger world states). To address the increased computational complexity in such challenging environments, we propose to develop efficient sampling techniques with data-driven evaluation functions and novel search heuristics to augment the Monte-Carlo search.

Lastly, our findings are limited to short-horizon interactions, as users only played two rounds of the game with each agent. To address this limitation, our proposed approach needs to be investigated in longitudinal interactions, where robots must anticipate and adapt to changes in user behavior or preferences over time.

# 8 CONCLUSION

In this work, we propose an online Bayesian approach, **Bayes-POMCP**, to optimize performance in mixed-initiative human-robot teams when both agents are suboptimal. Our focus is on learning a robot policy for effective user intervention. We design a novel domain inspired by USAR missions and conduct two user studies. Results from our first study indicate that robot interventions can improve team performance while recognizing diverse user preferences for different intervention styles. Next, we demonstrate that **Bayes-POMCP** can significantly improve team performance, compared to our baselines, both for different simulated human models and real users. Users also rated the **Bayes-POMCP** policy favorably with respect to the subjective metrics such as trust and likeability in our second study. In future work, we aim to further assess our algorithm for long-horizon interactions and extend it beyond grid-world domains to real-world human-robot collaboration tasks.

# ACKNOWLEDGMENTS

This work was sponsored by a gift from Konica Minolta and the National Institutes of Health (NIH) under Grant 1RO1HL157457.

## REFERENCES


[1] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. 2017. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence* 242 (2017), 132–171.

[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1 (2009), 71–81.

[3] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.

[4] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *arXiv preprint arXiv:1606.01540* (2016).

[5] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-AI coordination. *Advances in neural information processing systems* 32 (2019).

[6] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 307–315.

[7] Manolis Chiou, Nick Hawes, and Rustam Stolkin. 2021. Mixed-Initiative variable autonomy for remotely operated mobile robots. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 4 (2021), 1–34.

[8] Vanessa Evers, Henriette Cramer, Maarten van Someren, and Bob Wielinga. 2010. Interacting with adaptive systems. *Interactive collaborative information systems* (2010), 299–325.

[9] Douglas A Few, David J Bruemmer, and Miles C Walton. 2006. Improved human-robot teaming through facilitated initiative. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 171–176.

[10] Lewis R. Goldberg. 1992. The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment* 4, 1 (1992), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

[11] Matthew Gombolay, Anna Bair, Cindy Huang, and Julie Shah. 2017. Computational design of mixed-initiative human–robot teaming that considers human factors: situational awareness, workload, and workflow preferences. *The International journal of robotics research* 36, 5-7 (2017), 597–617.

[12] Matthew C Gombolay, Reymundo A Gutierrez, Shanelle G Clarke, Giancarlo F Sturla, and Julie A Shah. 2015. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots* 39 (2015), 293–312.

[13] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

[14] Joey Hong, Anca Dragan, and Sergey Levine. 2023. Learning to Influence Human Behavior with Offline Reinforcement Learning. *arXiv preprint arXiv:2303.02265* (2023).

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

[16] J Isaacs, Kevin Knoedler, Andrew Herdering, Mishell Beylik, and Hugo Quintero. 2022. Teleoperation for Urban Search and Rescue Applications. *Field Robotics* 2 (2022), 1177–1190.

[17] Hong Jun Jeon, Dylan P Losey, and Dorsa Sadigh. 2020. Shared autonomy with learned latent actions. *arXiv preprint arXiv:2005.03210* (2020).

[18] Shu Jiang and Ronald C Arkin. 2015. Mixed-initiative human-robot interaction: definition, taxonomy, and survey. In *2015 IEEE International conference on systems, man, and cybernetics*. IEEE, 954–961.

[19] Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. 2017. Uncertainty-aware reinforcement learning for collision avoidance. *arXiv preprint arXiv:1702.01182* (2017).

[20] Sammie Katt, Frans A Oliehoek, and Christopher Amato. 2017. Learning in POMDPs with Monte Carlo tree search. In *International Conference on Machine Learning*. PMLR, 1819–1827.

[21] Jill Kickul and George Neuman. 2000. Emergent leadership behaviors: The function of personality and cognitive ability in determining teamwork performance and KSAs. *Journal of Business and Psychology* 15 (2000), 27–51.

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

[23] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P. Losey, and Dorsa Sadigh. 2020. When Humans Aren't Optimal: Robots That Collaborate with Risk-Aware Humans. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/3319502.3374832

[24] Mikko Lauri, David Hsu, and Joni Pajarinen. 2022. Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics* 39, 1 (2022), 21–40.

[25] Joshua Lee, Jeffrey Fong, Bing Cai Kok, and Harold Soh. 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)*. IEEE, 6296–6303.

[26] Jin Joo Lee, Fei Sha, and Cynthia Breazeal. 2019. A Bayesian theory of mind approach to nonverbal communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 487–496.

[27] Xingzhou Lou, Jiaxian Guo, Junge Zhang, Jun Wang, Kaiqi Huang, and Yali Du. 2023. PECAN: Leveraging Policy Ensemble for Context-Aware Zero-Shot Human-AI Coordination. *arXiv preprint arXiv:2301.06387* (2023).

[28] B.M. Muir and B.M. Muir. 1989. *Operators' Trust in and Use of Automatic Controllers in a Supervisory Process Control Task*. University of Toronto. https://books.google.com/books?id=T94NSwAACAAJ

[29] Amal Nanavati, Christoforos I Mavrogiannis, Kevin Weatherwax, Leila Takayama, Maya Cakmak, and Siddhartha S Srinivasa. 2021. Modeling Human Helpfulness with Individual and Contextual Factors for Robot Planning.. In *Robotics: Science and Systems*.

[30] Manisha Natarajan and Matthew Gombolay. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 33–42.

[31] Manisha Natarajan, Esmaeil Seraj, Batuhan Altundas, Rohan Paleja, Sean Ye, Letian Chen, Reed Jensen, Kimberlee Chestnut Chang, and Matthew Gombolay. 2023. Human-Robot Teaming: Grand Challenges. *Current Robotics Reports* (2023), 1–20.

[32] Brenda Ng, Kofi Boakye, Carol Meyers, and Andrew Wang. 2012. Bayes-adaptive interactive POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 1408–1414.

[33] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems* 34 (2021), 610–623.

[34] Stefania Pellegrinelli, Henny Admoni, Shervin Javdani, and Siddhartha Srinivasa. 2016. Human-robot shared workspace collaboration via hindsight optimization. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 831–838.

[35] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*. PMLR, 4218–4227.

[36] Isabel Raemdonck and Jan-Willem Strijbos. 2013. Feedback perceptions and attribution by secretarial employees: Effects of feedback-content and sender characteristics. *European Journal of Training and Development* 37, 1 (2013), 24–48.

[37] Aditi Ramachandran, Sarah Strohkorb Sebo, and Brian Scassellati. 2019. Personalized robot tutoring using the assistive tutor pOMDP (AT-POMDP). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8050–8057.

[38] Aniketh Ramesh, Christian Alexander Braun, Tianshu Ruan, Simon Rothfuß, Sören Hohmann, Rustam Stolkin, and Manolis Chiou. 2023. Experimental Evaluation of Model Predictive Mixed-Initiative Variable Autonomy Systems Applied to Human-Robot Teams. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 5291–5298.

[39] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. 2007. Bayes-adaptive pomdps. *Advances in neural information processing systems* 20 (2007).

[40] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. 2016. Planning for autonomous cars that leverage effects on human actions.. In *Robotics: Science and systems*, Vol. 2. Ann Arbor, MI, USA, 1–9.

[41] Esmaeil Seraj, Jerry Yuyang Xiong, Mariah L Schrum, and Matthew Gombolay. 2023. Mixed-Initiative Multiagent Apprenticeship Learning for Human Training of Robot Teams. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[42] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[43] David Silver and Joel Veness. 2010. Monte-Carlo planning in large POMDPs. *Advances in neural information processing systems* 23 (2010).

[44] Herbert A Simon. 1972. Theories of bounded rationality, decision and organization. *CBR a. R. Radner. Amsterdam, NorthHolland* (1972).

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

[46] Ning Wang, David V Pynadath, and Susan G Hill. 2016. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*. 997–1005.

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