Offline Risk-sensitive RL with Partial Observability to Enhance Performance in Human-Robot Teaming

Giorgio Angelotti
ANITI, Fédération ENAC
ISAE-SUPAERO ONERA,
Université de Toulouse, France
giorgio.angelotti@isae-supaero.fr

Caroline P. C. Chanel
ANITI, Fédération ENAC
ISAE-SUPAERO ONERA,
Université de Toulouse, France
caroline.chanel@isae-supaero.fr

Adam Henrique Moreira Pinto
Fédération ENAC
ISAE-SUPAERO ONERA,
Université de Toulouse, France
adam.moreira-pinto@isae-supaero.fr

Christophe Lounis
Fédération ENAC
ISAE-SUPAERO ONERA,
Université de Toulouse, France
christophe.lounis@isae-supaero.fr

Corentin Chauffaut
Fédération ENAC
ISAE-SUPAERO ONERA,
Université de Toulouse, France
corentin.chauffaut@isae-supaero.fr

Nicolas Drougard
ANITI, Fédération ENAC
ISAE-SUPAERO ONERA,
Université de Toulouse, France
nicolas.drougard@isae-supaero.fr

ABSTRACT

The integration of physiological computing into mixed-initiative human-robot interaction systems offers valuable advantages in autonomous task allocation by incorporating real-time features as human state observations into the decision-making system. This approach may alleviate the cognitive load on human operators by intelligently allocating mission tasks between agents. Nevertheless, accommodating a diverse pool of human participants with varying physiological and behavioral measurements presents a substantial challenge. To address this, resorting to a probabilistic framework becomes necessary, given the inherent uncertainty and partial observability on the human's state. Recent research suggests to learn a Partially Observable Markov Decision Process (POMDP) model from a data set of previously collected experiences that can be solved using Offline Reinforcement Learning (ORL) methods. In the present work, we not only highlight the potential of partially observable representations and physiological measurements to improve human operator state estimation and performance, but also enhance the overall mission effectiveness of a human-robot team. Importantly, as the fixed data set may not contain enough information to fully represent complex stochastic processes, we propose a method to incorporate model uncertainty, thus enabling risk-sensitive sequential decision-making. Experiments were conducted with a group of twenty-six human participants within a simulated robot teleoperation environment, yielding empirical evidence of the method's efficacy. The obtained adaptive task allocation policy led to statistically significant higher scores than the one that was used to collect the data set, allowing for generalization across diverse participants also taking into account risk-sensitive metrics.

KEYWORDS

Human-Robot Interaction; Partial Observability; Offline RL; Risk-sensitive RL; Physiological Computing

1 INTRODUCTION

Over the past few years, Artificial Intelligence (AI) technologies have become widespread, playing a crucial role in automation across various industries and improving efficiency and productivity [24]. Hand in hand with the diffusion of AI, the need for regulations has been added to the agendas of law- and policy-makers. For instance, the recent European Union (EU) AI Act [17] promotes responsible AI deployment, emphasizing the need for robustness, interpretability, and human supervision, particularly relevant in Reinforcement Learning (RL). The integration of humans in the workflow of Machine Learning (ML) models has been conceived in multiple ways [45]. This field, called Human-In-The-Loop (HITL) ML [45], has been deemed necessary across several applications that require high-quality output and partial or total human supervision.

In this work, we focus on a specific niche of HITL-ML: HITL-RL and in particular HITL Offline RL (HITL-ORL). Offline RL can be seen as a completely separate branch of the RL community, aiming to obtaining an optimal control policy for an agent in an offline setting using a fixed batch of demonstrations. It is particularly well-suited to settings where the AI agent interacts with humans. In these scenarios, running parallelized and massive amount of simulations to train the agents is difficult or infeasible, since data collection involving human beings can be expensive, time-consuming, and possibly dangerous. Note that in the HITL-ORL context we consider, learning is not supervised by the human in the loop [2]. On the contrary, the human presence (*i.e.* the human included in the system) [47, 55] provides additional uncertain and partial observable features - as some relevant information about the human is not directly observable [53] - and poses significant difficulties for learning a policy with ORL methods. We specifically consider the context where ORL is used to obtain a control policy to drive the allocation of tasks between agents [27, 31, 54]. The control system

we aim to obtain should compute adaptive control policies, taking into account not only mission-related markers but also the state and behavior of the human operator.

Partially Observable Markov Decision Processes (POMDPs) have been used to make decisions in Human-Robot Interaction (HRI) [22] to deal with environment uncertainty and partial observability [20, 47, 48]. POMDPs allow for principled decision-making when human behavior is involved, including scenarios that involve multimodal observations (speech, eye gaze, and pointing gestures) [29, 46–48, 52, 55]. POMDP-based approaches are applied in various human-machine interaction cases, including robotics [38, 47, 48], medical diagnosis [59], and assisting students [56]. POMDP models in HRI are typically specified by experts [20], either fully or partially. Some research explores the possibility of robots learning parts of the POMDP model from user interactions, even in cases with unknown reward functions or observation models [47, 55]. However, as far as we know, no work has addressed in a unified manner the challenge of learning POMDP models from a fixed and limited data set, and the challenge of computing robust policies to the resulting (possibly inaccurate) models.

As previously stated, ORL aims to address the particular case when AI agents interact with the environment in contexts where trial-and-error learning could result in disastrous consequences [39, 50]. In ORL, the learning phase leverages a pre-collected data set. As discussed, the presence of humans, as part of the system, impacts the initial data set's informativeness in several ways: increased environmental stochasticity due to unpredictable behavior, limited data set size, and augmented biases. Addressing these challenges necessitates advanced ORL algorithms, such as [33, 40, 58?]. Interestingly, these model-based or model-free algorithms mainly focus on computing risk-sensitive or robust policies that could handle model uncertainties. Unfortunately, the state-of-the-art ORL algorithms [6, 33, 40, 49, 58?] were not designed to work with partially observable environments. Recently, Janner et al. [30] and Chen et al. [12] propose Transformer-based Deep Neural Network architectures that also deal with partial observable scenarios, but most Transformer-based methods do not perform at their best when confronted with limited data sets [57].

On top of this, which algorithm should a practitioner use among the many available ones? Interestingly, Angelotti et al. [4] proposed Exploitation vs Caution (EvC), a method for offline risk-sensitive policy selection in low-dimensional Markov Decision Processes (MDPs) that resorts to a Bayesian estimate of model uncertainty starting from a fixed and restricted set of previously collected experiences. EvC selects the most performing policy in a set of candidate ones by computing offline policy evaluation in a Monte Carlo fashion. By exploiting a Bayesian representation of model uncertainty, several models are sampled from the Bayesian posterior and every policy in the candidate set is evaluated on those models until the estimate of a quantile of the distribution of performance (according to model uncertainty) falls within a desired confidence interval. Then, the best policy according to a risk-sensitive metric is selected. Although this method does not scale for high-dimensional decision processes, it seems promising to adapt it to select risk-sensitive policies for data-driven POMDPs, thereby considering model uncertainty and including partial observability, in a data-frugal regime. In this sense, we address in a unified way the challenge of learning

a POMDP model and of computing robust policies for the resulting (possibly inaccurate) model. In this exciting context, the contributions of the present work are the following:

- We propose a methodology to approximate a POMDP model including a Bayesian representation of model uncertainty, and we implement it through a problem-specific pipeline representing a Human-Robot Interaction study case. In detail, the starting limited data set is the result of eighteen human operators who interacted with the robotic system in laboratory settings using a random interaction policy. Using this data set, the Human-Robot system is modeled as a POMDP, achieving an interpretable state space representation and a Bayesian representation of model uncertainty.
- We extend the EvC policy selection method to consider model uncertainty in partially observable domains. To do this, we first compute different policies for the data-driven POMDP model. Then, by resorting to the Bayesian formalism and Monte Carlo sampling, we extend the EvC method to handle partial observable domains and to select the safest policy according to a risk-sensitive measure and POMDP model uncertainty.
- We compare the robustness and the performance of the obtained robust POMDP policy with that used to collect the original data set, among others. This comparison is done by performing a novel set of experiments involving twenty-six human operators in laboratory facilities. The obtained robust POMDP policy led to statistically significant higher scores than the one that was used to collect the data set, allowing for generalization across diverse participants.

The paper is organized as follows: in the next section, we briefly review the technical background and related works. Then, in Section 3, we present the environment used for data collection and evaluation. In Section 4, the applied methodology is described, and in Section 5, we present our validation experiments and results. The paper concludes with a discussion and future work perspectives.

2 BACKGROUND AND RELATED WORKS

(PO)MDP Framework. Markov Decision Processes (MDPs) are a general mathematical framework that can describe the discrete-time progression of a Markovian stochastic process, while the actions of an agent can impact its evolution [44]. Partially Observable Markov Decision Processes (POMDPs) are an extension of the MDP framework in which the agent has access only to partial observations of the state of the environment [36]. In both cases, the agent receives a reward signal $r_t \in \mathbb{R}$ from the environment at each time step. Note that a Hidden Markov Model (HMM) can be seen as a particular POMDP case in which the state transition function does not depend on the actions and the reward function is not defined. Usually, solving a (PO)MDP means computing a policy, *i.e.* a function or a set of rules, that allows the agent to take actions to maximize the discounted expected cumulative reward obtained along a possible trajectory, $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $0 < \gamma < 1$ is the discount factor. The agent can compute such a policy by *solving* a (PO)MDP. However, solving a POMDP for an infinite time horizon is undecidable [43], and usually only approximate solutions can be obtained in reasonable time [36].

Offline POMDP model learning and solving. When partial observability enters the scene, learning a model from demonstrations becomes an exceedingly complicated task. Nevertheless, offline POMDP learning is of great importance as a representative model of the system dynamics might be impossible to be written by hand, especially when there are humans in the loop. An important issue in offline POMDP learning is intrinsic and rooted at the foundation of any possible learning procedure: which representation is the most appropriate? Remember that in a POMDP, the agent has only access to *observations*, while the system evolves over time by transitioning across changing, partially observable states. To implement any learning algorithm, one must first deal with the choice of how to represent the (hidden) state space. A second main issue is model uncertainty. For instance, in the case of a discrete MDP, it is relatively straightforward to include uncertainty about the model estimate within a Dirichlet distribution, initialized with the frequency of transitions in the data set. However, such convenient modeling is not available when dealing with POMDPs. Moreover, one should account for not only the uncertainty about the transition function but also the observation function. Offline POMDP learning is a problem far more difficult than offline MDP learning.

Offline POMDP learning has been investigated by the research community, but to the best of our knowledge, no one has managed to develop a general approach. The majority of existing methods that learn a POMDP offline to address planning for robotics define the representation with expert guidance. In the works from [5, 23, 56], the POMDP representation is provided by an oracle, and only transition and observation functions are learned. The work in [8] assumes that the states are described by discrete variables and the observations can be just a subset of the latter. In the context where the histories composing the data set have been generated by an unknown POMDP, the work in [19] shows that choosing a discrete state space with a lower dimensionality than the original one can help reduce overfitting and hence yield relevant policies. Interestingly, the same work also advocates that reducing the discount factor during planning can further benefit the performance of the policy at the time of deployment. Since learning a POMDP model from data is already challenging, model uncertainty is a difficulty that very few approaches have attempted to include and address. The works in [5, 14] acknowledge that a learned trivial model might not be representative enough. For this purpose, they established a Bayesian framework from which models are extracted by a Bayesian prior. However, model uncertainty is then reduced in an online setting by allowing the agent to interact with an oracle.

In this context, and inspired by the Bayesian framework proposed in [5, 14], we believe that extending a method for offline risk-sensitive policy selection resorting to a Bayesian estimate of model uncertainty, namely Exploitation vs Caution [4], to POMDPs, could be useful for selecting the most risk-sensitive policy among those obtained using different solvers and/or varying hyperparameters - e.g. γ as suggested by François-Lavet et al. [19]. Indeed, in a (PO)MDP, $1 - \gamma$ can be interpreted as the time-step probability to exit from the modeled stochastic process [37]. Jiang et al. [32] shows that in the context of data-driven model learning there is an optimal discount factor $\gamma^* < \gamma < 1$ that should be used to compute a policy for eventual real-world deployment. In light of this, our

approach aims to use EvC and select a policy computed with a discount factor estimated to be closer to γ^*.

3 THE FIREFIGHTER ROBOT GAME

The environment under study consists of a serious game, which can be played at robot-isae.isae.fr, designed to generate deleterious cognitive states in human operators during the missions they carry out with an artificial agent, mainly due to multitasking, uncertainty, and time pressure. Such situations are known to generate stress and cognitive workload, thus impacting human agent performance [13, 15, 16]. Deleterious mental states affecting performance can be estimated using physiological computing [18, 53]. For instance, metrics such as the Heart Rate (HR) and the Heart Rate Variability (HRV) are known to be impacted by workload [26]. With this in mind, we believe that exploring human behavioral and physiological features in the construction and evaluation of human-system interaction policies is particularly promising. In this context, our goal is to use ORL to develop a robust policy for an interaction controller within a quasi-mixed initiative interaction system - the system adapts its behavior based on several mission parameters, including those related to the human operator.

Brief environment description. The Firefighter Robot Game (FRG) [9, 11, 15] is a scenario where a human teleoperates a robot to extinguish fires in a confined area. The forest has nine trees that can catch fire, and the human-robot team must put out as many fires as possible within 10 minutes. The robot has limited battery power, a water tank for extinguishing fires, and a thermometer to monitor its temperature. The robot must recharge at an energy supply zone and refill its water tank at a water pool. However, the pool's walls are susceptible to leaks and necessitate manual intervention for mending. Additionally, filling the pool requires precise control of the movement of a moving filling nozzle.

An automated controller manages the human-robot interaction, switching between automatic and manual modes and activating or deactivating alarm notifications. In automatic mode, the robot prioritizes battery recharging and embedded water tank refilling. It navigates to the energy supplier or water pool when resources are low. However, the human operator is required to manage, repair, and refill the water pool. When battery and tank levels are sufficient, the robot finds the shortest path to the nearest burning tree and extinguishes the fire. In manual mode, the human operator remotely controls all robot actions, including navigation, recharging, temperature monitoring, and water dispensing.

Data collection. Data were collected in a former study involving 18 participants [10]. These participants underwent four missions, preceded by electrocardiogram (ECG) setup and eye-tracking calibration. Each mission was preceded by a rest period and followed by the filling of surveys and a break. During missions, the robot mode (automatic or manual) changed randomly every 10 seconds, as did the activation of the alarm system. Metrics such as robot parameters, mission states, operator actions, ECG and eye-tracking data were recorded. Artifacts were removed from live ECG data to compute the HR and HRV measurements. Rest data helped to compute normalized HR and HRV metrics for each mission and participant. Eye-tracking data consisted of the number and duration of

fixations in five areas of interest (see [10] for more details). Human actions on the interface, such as keystrokes and clicks, were also recorded. These characteristics were computed using 10-second intervals, and associated with the current robot's automation mode, and status of the alarm system. Henceforth, this data set will be denoted as $\mathcal{D}$. The success of a mission is quantified by the number of fires extinguished, overlooking other human-centric metrics.

4 RISK-SENSITIVE ORL WITH PARTIAL OBSERVABILITY

The Figure 1 illustrates the proposed pipeline. The data set $\mathcal{D}$ was used as input for ORL. We adopted a model-based approach, such as using a POMDP representation, for its superior interpretability. The POMDP's hidden states marked team performance, while the visible ones indicated robot autonomy mode and alarm status. We considered four actions: *put-manual-alarms-on, put-manual-alarms-off, put-automatic-alarms-on, put-automatic-alarms-off.* For the partial observable case, we considered an observation set identical to the state set as suggested by [8, 19]. Note that the observability concerning robot mode and alarm status is complete, although the observability concerning the current human-robot team performance is - by definition - partial. To learn the dynamics of such a model and the respective observation function, the methodology applied is hereafter described.

Data split. The mission score in $\mathcal{D}$, which represents the total number of extinguished fires throughout an entire mission, guided data splitting and classifier training to map high-dimensional observations (raw recorded data) into a low-dimensional variable, whose two possible entries are *performant* and *non-performant.* Specifically, the batch $\mathcal{D}$ consists of 72 mission recordings, each

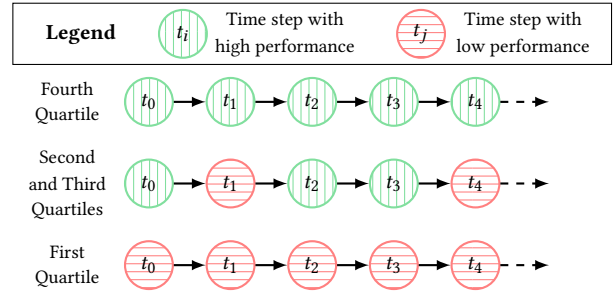**Figure 1: Illustration of the proposed methodology.**

Figure 2: Example of how the batch is split.

containing several feature vectors labeled by the mission's global score. However, to infer the system state variable concerning performance (including the mental engagement of a human operator), we need to evaluate a local measure of efficiency. To do this, we assume that during missions with a low score, the team exhibits constant non-performant behavior and, conversely, during missions with a very high score the team exhibits constant performant behavior. Hence, the time steps of the missions in the First Quartile (with respect to the final score) are labeled as *non-performant* (red in Fig. 2) and the ones in the top 25% (Fourth Quartile) as *performant* (green). The time steps in the missions in the second and third quartiles form the batch that will be used to learn the transition dynamics, as we assume that there might be frequent changes in performance status during missions with an average score.

Classifier training. Using the data set partitioned and labeled as mentioned above, we train four different classifiers. These classifiers provide an observation of the mission performance for the four possible configurations of the two remaining visible state variables, namely the robot autonomy mode (autonomous/manual) and the alarm system status (on/off). Each trained classifier translates the features, related to time steps with one of the four visible state configurations, into a (noisy) observation of the hidden variable, namely mission performance.

We opted to train four Extra Tree Classifiers as indicated by Geurts et al. [21]. We preferred Extra Tree Classifiers to Random Forests because the first produce results with reduced variance, even though with a slight increase in bias. This decision fits well with our objective, given that we aim for classifiers that can effectively generalize across various participants who may present a varied distribution in their measurements. The classifiers are trained using a 10-fold Group (participant-wise) Shuffle Split cross-validation approach to select the best hyperparameters with respect to the balanced accuracy metric, initially using a Random Search and then with a Grid Search method. The validation set contains 20% of the data from the entire training set, and both sets include only missions carried out by different groups of participants. In doing so, we obtain the hyperparameters of the Extra Tree Classifiers that, on average, minimize the generalization error on the validation set.

Observation Function and Bayesian prior. The EvC paradigm requires a Bayesian representation of model uncertainty. We note that the diagonal elements of the confusion matrix associated with each classifier enumerate the number of samples correctly classified by
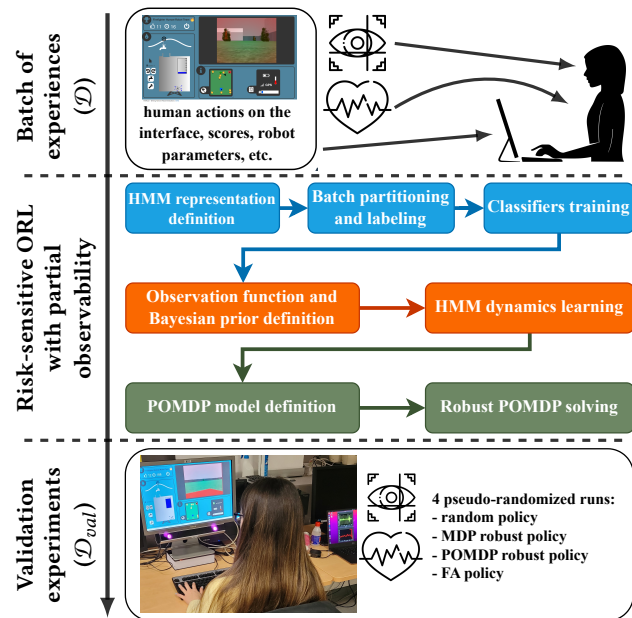
the respective algorithm. Conversely, the off-diagonal elements represent the number of samples classified as pertaining to a particular class (as determined by the column index), while their actual label belongs to a different class (as determined by the row index). Hence, we propose using the confusion matrices of the classifiers to compute Bayesian Dirichlet posteriors for an HMM observation function. Moreover, normalizing a confusion matrix by row yields another matrix whose elements are $O_{ij} \approx \Pr\left(O_t = o_j | S_t = s_i\right)$. These last row-normalized confusion matrices could also serve as the observation function for defining an HMM process. By defining an observation function - either by sampling one from the posteriors or simply by row-normalizing the prior - we can learn the dynamics of an HMM via Expectation Maximization (EM) [7]. The latter will be used to learn the dynamics of a POMDP model. Indeed, the uniform randomness of the data collection policy allows the HMM to be transformed into a POMDP.

POMDP model definition. We define the POMDP model, called the *trivial POMDP*, that we will use both to obtain the robust policy and to deploy the policy at execution time during experiments with new human participants. Indeed, we will need this model to update the belief in real-time. The trivial POMDP is obtained by using the row-normalized confusion matrices as observation function and the subsequent transformed HMM dynamics inferred via EM. As reward function, we use the average number of fires extinguished by the human-robot team in the original data-set for each state.

Robust POMDP solving. Inspired by the works in [19, 32], the trivial POMDP was solved using SARSOP [36] with different discount factors, producing diverse policies (Lines 2-4, Algorithm 1). We then adapted the EvC algorithm for risk-aware policy selection in the POMDP context. Non-normalized confusion matrices updated Dirichlet priors expressing observation model uncertainty. Diverse observation functions were sampled from these posteriors (Line 6), yielding different HMM dynamics (Line 7) and POMDPs (Line 8). Candidate policies were evaluated with these POMDPs (Lines 9-14), with beliefs updated based on the solved model. The Value-at-Risk [51] at risk level q, *i.e.* the value of the q-order quantile, set to 0.5, guided robust policy selection (Line 15). As a limitation of the approach, we do not continue sampling models until the estimates of the quantile for each policy falls within a confidence interval, as in [4], due to computational time constraints. Nevertheless, although lacking theoretical guarantees, this method seems to produce decent results in the tested scenario.

5 VALIDATION EXPERIMENTS

By performing validation experiments in the same environment (the Firefigther Robot Game) with different participants than those gathered during the initial data collection stage, the robustness and transferability of the policies can be fairly assessed. Preliminary power analysis was performed, requiring at least 30 participants to perform a significant Friedman test with four groups at 0.95 power (four policies will be compared). Participants were incrementally recruited through electronic mail announcements and oral advertising, and recruitment stopped when the experiment's estimated power reached 0.95. In the end, 26 participants (mean age 28.6, sd. 5.7; 7 females) took part in the experiments that were approved by

---

Algorithm 1: Robust POMDP solving and Risk-sensitive Policy Selection

---

Input: Trivial POMDP, Γ set of discount factors, $\mathcal{D}$ batch of trajectories, N_M number of models to sample/learn, N_E number of histories per model, q quantile's order

1 **Initialization:** $M \leftarrow \varnothing$ (empty list), $\Pi \leftarrow \varnothing$ (empty list)

2 **forall** γ *in* Γ **do**

3     $\pi \leftarrow$ Solve(Trivial POMDP, γ)

4   Append π to Π

5 **for** i *from* 1 *to* N_M **do**

6   Sample observation functions from the Dirichlet posterior distributions

7   Learn transition functions using EM(obs. functions, $\mathcal{D}$)

8   Append new POMDP model to M

9 **forall** π *in* Π **do**

10     $G_\pi \leftarrow \varnothing$ (empty list)

11   **forall** *pomdp in* M **do**

12   **for** i *from* 1 *to* N_E **do**

13           $R \leftarrow$ Total reward of a generated trajectory

14   Append R to G_π

15 $\pi^* \leftarrow \arg\max_{\pi \in \Pi} \text{VaR}_q[G_\pi]$

16 **return** π^*

---

our local ethics committee (CER-2018-070). From now on, the data set including the validation experiments will be referred as $\mathcal{D}_{val}$. Fig. 1 shows a participant taking part in the experiment.

5.1 Protocol

The introductory phase of the experimental procedure included collecting informed consent, responsibility signatures, providing explanations, calibrating physiological measurement devices, preliminary training, and administering the Karolinska Sleepiness Scale (KSS) [1] questionnaire. KSS is generally used to measure the participants' subjective fatigue; a higher post-experiment KSS score can indicate significant drowsiness. The core procedure of the experiments involved four missions in the FRG environment, each applying a distinct interaction control policy. These policies were pseudo-randomized among participants. Each run commenced with a baseline recording of cardiac activity (see Fig. 5), then proceeded to mission execution. Average HR and HRV during a 1-minute resting period were used as baselines to normalize subsequent HR and HRV values. Artifacts in HR and HRV measurements were removed in real-time. Participants then completed the NASA-TLX questionnaire [25], providing feedback on the perceived effort generated by the mission. Subsequently, they filled out a questionnaire inspired by the work in [28], to provide feedback on the fluency of the current adaptive policy before taking a 2-minute break.

The experiments were conducted with four interaction control policies: robust POMDP solution, robust MDP solution, a random policy [10] (the very same data collection policy), and a policy where the robot mode was fixed to automatic with alarms. These conditions were pseudo-randomized among participants. The robust POMDP policy was computed as detailed in Sec. 4 with
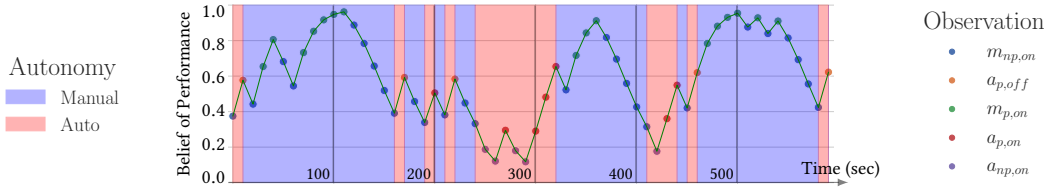
Figure 3: Computation of POMDP's marginalized belief of performance β_t along a mission run using the POMDP policy. The control system switches the autonomy mode of the robot from manual to automatic (auto), and vice-versa. Notice how the belief is different from the immediate observation, as multiple subsequent *non-performant* observations can be necessary for the system to deem the state necessary to put the robot in automatic mode, e.g. from $t \approx 100s$ to $t \approx 150s$.

$\gamma = \{0.7, 0.8, 0.9, 0.97, 0.98, 0.99\}$ and sampling $N_M = 10000$ models. For each sampled model, each policy was deployed during $N_E = 10000$ histories. The policy corresponding to $\gamma = 0.98$ was selected because it had the highest estimated Value-at-Risk at risk-level 0.5. The MDP robust policy was obtained as follows: once the classifiers had been trained (third blue box in Fig. 1), instead of learning an observation function and an HMM dynamics, the classifiers were directly used to perform dimensionality reduction on the part of the data set split to learn the dynamics, converting hence the multi-dimensional and multi-modal time series of observations into time series of discrete states defining an MDP. On this transformed set of discrete MDP trajectories, several data-driven risk-sensitive baselines - e.g. soft-robust optimization [40] - were executed, producing a set of candidate policies. Due to the low state-action space dimensionality, the EvC algorithm drove risk-sensitive policy selection, using the same criteria as for the POMDP case. Surprisingly, the robust MDP policy resulted in a fully manual mode with alarms, where the human operator was confronted to the multitasking paradigm during the whole mission. In contrast, with the Fixed Automatic (FA) mode interaction control policy, the human faces an easier task with respect to the other interaction conditions and is less prone to deleterious mental states. In practice, we would like to use the automatic mode only when the human operator is estimated unable to drive the robot, and not during the whole mission. Hence, although the results obtained with this policy are presented for discussion, its comparison with the policies calculated by our method is outside the scope of this work. After the four iterations, participants completed the KSS questionnaire again. The entire experiment lasted approximately 1 hour and 30 minutes per participant, totaling 39 hours.

5.2 Results and Analysis

In this subsection, we elaborate on the results. The subsequent analysis will cover robustness, data informativeness, data scarcity, inter-subject variability, and the explainability of the implemented approach. We will not discuss the comparison with the FA policy because: (i) it was not learned from data using an ORL approach; (ii) it served to establish the best possible average score of the environment, in a focused sub-task distinct from other conditions; (iii) in critical missions, such a policy is inadvisable due to the necessity of operator vigilance and control (for responsibility issues).

POMDP policy. We start by analyzing qualitatively the POMDP risk-sensitive policy. With this policy, the controller changes the

autonomy mode of the robot from manual mode to automatic mode when the estimated marginalized belief of performance β_t falls below a given threshold - i.e. the system's belief is that the state is *non-performant*. This policy, obtained with the paradigm previously outlined, is reasonable, since in $\mathcal{D}$, the time steps associated with the manual mode and performant operators were observed to be related to a higher score, while time steps with a non-performant operator had the highest reward when the robot was in automatic mode. For illustration purposes, an online β_t computation during a mission run using the robust POMDP policy, along with observations received and deployed actions, is displayed in Figure 3.

Comparing the distribution of scores. Descriptive statistics for the results can be found in Table 1. We started by comparing the score distribution obtained with the random interaction policy evaluated in data set $\mathcal{D}_{val}$ with the data in the starting batch $\mathcal{D}$, that was also collected using a random policy (first line of Table 1). The Mann-Whitney U test was conducted to compare the two independent groups of scores. The test revealed a significant difference between the groups ($U = 686.0$, p-value < 0.05), rejecting the null hypothesis of no difference between the groups at the 5% significance level. We can conclude that $\mathcal{D}$ and $\mathcal{D}_{val}$, even when the interacting policy is the same, do not follow the same distribution. This finding underscores the challenge of our task, as we had to work with a limited data set $\mathcal{D}$ and huge inter- and intra- subject variability.

Robustness. To assess whether applying an adaptive control policy leads to more robust results than deploying the policy used

Table 1: Descriptive statistics of the scores by interaction control policy (random or learned) and data set. $\mathcal{D}$ is the data set initially collected. $\mathcal{D}_{val}$ is the data set including the results of the experimental validation of the HITL-ORL pipeline. Best value per metric is displayed in bold. FA states for Fixed-Automatic control policy, leading to a single-task mission.

Policy	mean	std	min	25%	Median	75%	max
Random ($\mathcal{D}$)	22.1	10.0	1.0	12.8	26.5	29.3	36.0
Random ($\mathcal{D}_{val}$)	17.9	9.6	4.0	8.3	20.0	26.8	32.0
MDP ($\mathcal{D}_{val}$)	22.7	7.6	6.0	18.0	22.5	**28.8**	**39.0**
POMDP ($\mathcal{D}_{val}$)	**23.4**	**5.7**	**14.0**	18.5	**24.0**	27.5	37.0
FA ($\mathcal{D}_{val}$)	25.6	5.0	9.0	25.3	27.0	28.0	32.0

to collect the data, we performed a Friedman's statistical test to analyze the data of scores during the missions in $\mathcal{D}_{val}$. The results revealed a significant difference among the policies ($\chi^2(3) = 13.07$, $p < 0.01$). Post-hoc analyses using the Wilcoxon signed-rank test with a Hommel correction indicated that the Random policy (Median = 20) yielded significantly lower scores compared to the MDP policy (Median = 22.5, $p < 0.05$), and the POMDP policy (Median = 24, $p < 0.05$). Hence, it is convenient to deploy a policy learned with the proposed HITL-ORL method. Despite the better results of the POMDP policy for risk-sensitive metrics in Table 1 (minimum score more than double the minimum one achieved with the MDP policy, and a lower standard deviation than the random/MDP policies), no statistically significant difference was observed between the MDP and the POMDP robust policies. The MDP policy maintains the manual mode, so the mission is therefore entirely multi-tasking, allowing users who are comfortable with this context to perform better, but making others perform very poorly. With the POMDP policy, the robot's mode depends on user state estimation (Fig. 3).

Validity of the POMDP models. To assess the POMDP models we calculate the POMDP belief for each time step, β_t, that represents the probability that during that time step the human-robot team is performant. The said β_t is computed for each mission, including those run with a policy other than the POMDP one (obtained with EvC using SARSOP). If the model is accurate, we anticipate a high average belief of performance, $\overline{\beta} = \frac{1}{T}\sum_{t=1}^{T}\beta_t$ in missions with high scores, and *vice-versa*. Thus, we compute the correlation between $\overline{\beta}$ and the score for each mission. The Spearman's rank correlation reveals a significant positive correlation between $\overline{\beta}$ and the global mission score ($\rho(\overline{\beta}) = 0.325$, $N(\overline{\beta}) = 153$, $p(\overline{\beta})$-value < 0.001). In Figure 4, we illustrate the relationship between $\overline{\beta}$ and the mission score, along with a fitted (monotonic) fifth-order polynomial. A positive correlation suggests alignment of the POMDP with actual events, despite classifiers with balanced accuracies around 0.67. The Spearman's correlation coefficient only identifies monotonic correlations, without probing variable interdependence. To evaluate dependence between $\overline{\beta}$ and the score, we used the Randomized Dependence Coefficient (RDC) [41], using suggested hyperparameters from the reference. The resulting RDC was 0.348, almost identical to the Spearman's coefficient, suggesting a likely absence of non-monotonic dependence between the variables.

Analysis of the used classifiers and Explainability. The confusion matrices for the eight classifiers, four computed on the test set (in $\mathcal{D}$) and four on the data set obtained from the experimental results ($\mathcal{D}_{val}$) and split according to the same criterium were analyzed with a two-proportion Welch's t-test (see Supplementary Material [3]). As the p-value is less than or equal to 0.05 for almost all states, we can reject the null hypothesis that the classification outcomes have equivalent proportions. As a result, at the very least, the classifiers cannot be used for an *accurate* representation of the process we aimed to model with both an MDP and/or a POMDP. This effect can be due to data scarcity and insufficient data informativeness, and also to an intense inter-subject variability both in the training set in $\mathcal{D}$ and in the set of experimental results $\mathcal{D}_{val}$.

To study which features were important for the classifier we used SHAP [42] (see Supplementary Material [3]). The normalized
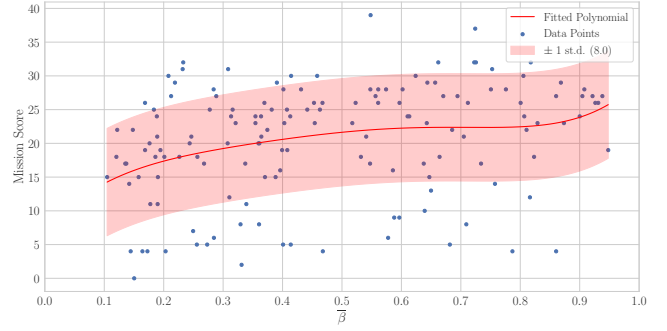

Figure 4: Scatter plot displaying the relationship between the average belief of human-robot performance during a mission (x-axis) and the mission score (y-axis). A fifth-order (monotonic) polynomial is fitted to show the positive correlation between $\overline{\beta}$ and the score. The shaded area represents plus or minus one standard deviation of the residuals of the fit.

HRV (a cardiac feature) in the 10-second time window is among the two most important features used for performance classification. Since the HRV is obtained after the standardization procedure of subtracting the resting baseline from the live measurement, any anomaly or perturbation in the standardization can have a huge impact on the classification. The explainability analysis provides great insights into the usefulness of incorporating physiological computing for HITL systems. However, inter-subject variability, and improper processing may lead to unwanted results. The inter-subject variability for the HRV at rest before each mission in the batch data set is displayed in Figure 5. In spite of the normalization attempts, we observed notable discrepancies in the recorded resting HR and HRV baselines for the same individual across different missions (see Supplementary Material [3]). The discrepancies were even more pronounced for HRV measurements at rest: the disparity between the maximum and minimum measurements of HRV at rest per subject could be as much as eightfold (see Subject 6 in Fig5).

Subjective feedback from participants. A Wilcoxon signed-rank test using the *z-split* method was conducted to analyze the KSS
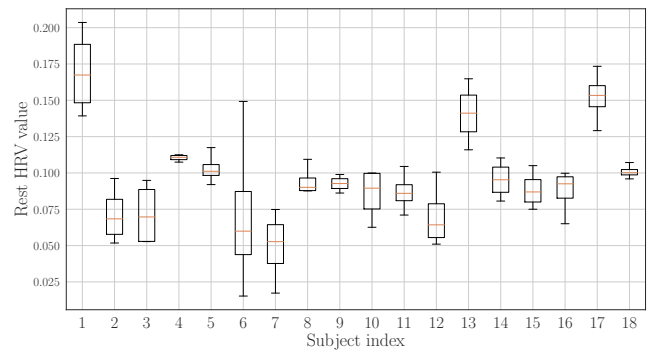

Figure 5: Rest HRV per subject before a mission. Since these values are used to normalize the features to be used by the ML model, variability could lead to a failure of the pipeline.

scores before and after the experimental procedure. The test indicated that the difference between the two sets of scores was not statistically significant ($z = 143.5$, $p = 0.41$). This means that participants did not report subjective fatigue after the experiments, suggesting they remained committed to the task. For the NASA-TLX questionnaire scores, a Friedman's test was performed to analyze the overall rating, revealing a significant difference among the policies ($\chi^2(3) = 36.98$, $p < 0.001$). Post-hoc analysis using the Wilcoxon signed-rank test with Hommel correction showed a significant difference between the ratings obtained when the controller followed the fixed-automatic (FA) policy (Median = 40.0) and the other policies with $p < 0.001$ (Random, Median = 61.2; MDP, Median = 71.0; POMDP, Median = 65.3). These findings indicate that the FA policy, as expected, is perceived as the less demanding policy; and that the impact on the subjective workload of adaptive policies, unfortunately, remains inconclusive. Surprisingly, the results of the Fluency survey showed the FA policy (Median = 5.2) was perceived as providing the most fluent interaction, while the MDP adaptive control policy is considered the least fluent (Median = 2.0). A Friedman's test was conducted to analyze the overall rating, revealing a significant difference among the policies ($\chi^2(3) = 32.72$, $p < 0.001$). Post-hoc analysis using the Wilcoxon signed-rank test with Hommel correction showed a significant difference between the ratings obtained for every pair of policies ($p < 0.01$), except for the pair made of the Random (Median = 4.0) and the POMDP (Median = 4.5) policies ($p = 0.62$). After discussion with participants, we speculate that FA policy (the easier experimental condition in which no adaptive interaction was used, as the robot was navigating autonomously all the time) favored some participants' experience. They reported that they could concentrate on the tank management tasks because they were confident in the robot's behavior, thus preferring such a constant task allocation policy. Despite this, we believe the findings on the random policy and adaptive POMDP policy comparison warrant further investigation to better understand the potential of the latter in enhancing human-robot interaction experience and performance (see Table 1).

6 CONCLUSIONS AND FUTURE WORK

In this work, we propose a methodology to approximate a POMDP model that includes a Bayesian representation of model uncertainty. We implement this through a specific use-case: the Firefighter Robot Game, a human-robot interaction proof-of-concept scenario. Physiological and behavioral features were integrated into the system, providing useful insights for calculating an interaction control policy, which determines the robot's autonomy during the mission. The control policies were obtained in an offline setting using only a previously collected set of human-robot interaction experiences. This data set comprised data from eighteen human operators who interacted with the system in laboratory facilities using a random interaction policy. The FRG is modeled as a POMDP, achieving an interpretable state space representation and a Bayesian representation of model uncertainty.

Given the limited size of the data set and the relatively small number of volunteers, we assumed that we did not possess enough information to capture the full stochastic description of the environment. Therefore, we resorted to the Bayesian formalism to estimate uncertainty over the set of possible models, taking inspiration from the EvC [4] method. More precisely, we extended the EvC policy selection method to consider model uncertainty in partially observable domains. To do this, we first computed a set of different policies for a data-driven trivial POMDP model using different discount factors (γ's). Then, using Bayesian formalism and Monte Carlo sampling, we exploited model uncertainty to select a policy from among the set of candidate policies, according to a risk-sensitive metric.

We compared the robustness and the performance of the obtained adaptive POMDP policy with that used to collect the original data set, among others. This comparison was made by performing a new set of experiments involving twenty-six human operators in laboratory facilities. The experiments showcased that the proposed paradigm led to significantly higher performance than that obtained with the data collection policy. Furthermore, our results suggest that considering partial observability is propaedeutic to more robust policies. Moreover, by scrutinizing our experimental results, we identified several obstacles to the application of ORL, particularly highlighting the opportunities and challenges that arise when a human operator is included in the environment. Data representativeness and scarcity were major concerns; proper statistical tests were necessary to individuate these issues. A lack of sufficient diversity in data sets hindered generalization and would have required appropriate diverse recruitment campaigns and possibly more data. Nevertheless, the statistical tests showed a positive correlation between the belief in the human-robot team's performance and the mission score.

Regarding the classifiers used to map high-dimensional raw features into a low-dimensional observation space, we note that they relied heavily on physiological markers like standardized HRV for classification. The approach used to normalize HR and HRV features by subtracting resting baselines revealed potential discrepancies. As classifiers may favor these features, results could be compromised. On one hand, implementing an efficient experimental protocol that considers physiological and behavioral measurements and their variability among participants remains an open question. On the other hand, addressing partial human state observability and implementing ORL methods to handle model uncertainty can mitigate the effects of poor data curation and scarcity, thereby improving generalization.

Perspectives. Future work could encompass modifying the FRG to allow the human operator to set the autonomy mode of the robot whenever desired, permitting thus to study the application of ORL and physiological computing to scenarios characterized by complete mixed-initiative human-robot interaction. Moreover, in the light of the current study, it might be appropriate to collect larger data sets before computing the risk-sensitive control policies with offline reinforcement learning, or to limit the analysis to a more homogeneous set of human operators in order to reduce inter-subject variability.

REFERENCES

[1] Torbjörn Åkerstedt and Mats Gillberg. 1990. Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience* 52, 1-2 (1990), 29–37.

[2] Sharath Chandra Akkaladevi, Matthias Plasch, Sriniwas Maddukuri, Christian Eitzinger, Andreas Pichler, and Bernhard Rinner. 2018. Toward an Interactive Reinforcement Based Learning Framework for Human Robot Collaborative Assembly Processes. *Frontiers in Robotics and AI* 5 (2018), 126.

[3] Giorgio Angelotti, Caroline P. C. Chanel, Adam H. M. Pinto, Christophe Lounis, Corentin Chauffaut, and Nicolas Drougard. 2024. Offline Risk-sensitive RL with Partial Observability to Enhance Performance in Human-Robot Teaming. arXiv:2402.05703 [cs.MA]

[4] Giorgio Angelotti, Nicolas Drougard, and Caroline Ponzoni Carvalho Chanel. 2023. An Offline Risk-aware Policy Selection Method for Bayesian Markov Decision Processes. (2023). arXiv:2105.13431.

[5] Amin Atrash and Joelle Pineau. 2010. A Bayesian method for learning POMDP observation parameters for robot interaction management systems. In *The POMDP practitioners workshop*.

[6] Bahram Behzadian, Reazul Hasan Russel, Marek Petrik, and Chin Pang Ho. 2021. Optimizing Percentile Criterion using Robust MDPs. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*. PMLR, 1009–1017.

[7] J. Bilmes et al. 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute* 4, 510 (1998), 126.

[8] Frank Broz, Illah Nourbakhsh, and Reid Simmons. 2011. Designing pomdp models of socially situated tasks. In *2011 RO-MAN*. IEEE, 39–46.

[9] Caroline PC Chanel, Raphaëlle N Roy, Nicolas Drougard, and Frédéric Dehais. 2020. Mixed-initiative human-automated agents teaming: towards a flexible cooperation framework. In *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*. Springer, 117–133.

[10] Caroline Ponzoni Carvalho Chanel, Raphaëlle N Roy, Frédéric Dehais, and Nicolas Drougard. 2020. Towards mixed-initiative human–robot interaction: Assessment of discriminative physiological and behavioral features for performance prediction. *Sensors* 20, 1 (2020), 296.

[11] Jack-Antoine Charles, Caroline Ponzoni Carvalho Chanel, Corentin Chauffaut, Pascal Chauvin, and Nicolas Drougard. 2018. Human-agent interaction model learning based on crowdsourcing. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. 20–28.

[12] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision Transformer: Reinforcement Learning via Sequence Modeling. In *Advances in Neural Information Processing Systems*, Vol. 34. 15084–15097.

[13] Frédéric Dehais, Alex Lafont, Raphaëlle Roy, and Stephen Fairclough. 2020. A neuroergonomics approach to mental workload, engagement and human performance. *Frontiers in neuroscience* 14 (2020), 268.

[14] Finale Doshi-Velez, Joelle Pineau, and Nicholas Roy. 2012. Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs. *Artificial Intelligence* 187 (2012), 115–132.

[15] Nicolas Drougard, Caroline Ponzoni Carvalho Ponzoni Chanel, Raphaëlle N Roy, and Frédéric Dehais. 2017. Mixed-initiative mission planning considering human operator state estimation based on physiological sensors. In *IROS-2017 workshop on Human-Robot Interaction in Collaborative Manufacturing Environments (HRI-CME)*. https://caris-mech.sites.olt.ubc.ca/files/2017/09/HRI-CME_2017_paper_6.pdf

[16] Nicolas Drougard, Raphaëlle N Roy, Sébastien Scannella, Frédéric Dehais, and Caroline Ponzoni Carvalho Chanel. 2018. Physiological Assessment of Engagement during HRI: Impact of Manual vs Automatic Mode. In *2nd International Neuroergonomics Conference*. 0.

[17] European Commission. 2020. White Paper on Artificial Intelligence - A European approach to excellence and trust. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. [Online; accessed 09-October-2023].

[18] Stephen H Fairclough. 2009. Fundamentals of physiological computing. *Interacting with computers* 21, 1-2 (2009), 133–145.

[19] Vincent François-Lavet, Guillaume Rabusseau, Joelle Pineau, Damien Ernst, and Raphael Fonteneau. 2019. On Overfitting and Asymptotic Bias in Batch Reinforcement Learning with Partial Observability. *Journal of Artificial Intelligence Research* 65 (2019), 1–30.

[20] Thibault Gateau, Caroline P Carvalho Chanel, Mai-Huy Le, and Frédéric Dehais. 2016. Considering human's non-deterministic behavior and his availability state when designing a collaborative human-robots system. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4391–4397.

[21] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (2006), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[22] Michael A. Goodrich, Alan C. Schultz, et al. 2008. Human–Robot Interaction: A Survey. *Foundations and Trends® in Human–Computer Interaction* 1, 3 (2008), 203–275.

[23] Nakul Gopalan and Stefanie Tellex. 2015. Modeling and Solving Human-Robot Collaborative Tasks Using POMDPs. In *RSS Workshop on Model Learning for Human-Robot Communication*, Vol. 32. 590–628.

[24] Wendy Hall and Jérôme Pesenti. 2017. Growing the artificial intelligence industry in the UK. (2017).

[25] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. 52 (1988), 139–183. https://doi.org/10.1016/s0166-4115(08)62386-9

[26] Jamison Heard, Caroline E Harriott, and Julie A Adams. 2018. A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems* 48, 5 (2018), 434–451.

[27] Marti A. Hearst, J. Allen, C. Guinn, and Eric Horvitz. 1999. Mixed-initiative interaction: Trends and controversies. *IEEE Intelligent Systems* 14, 5 (1999), 14–23.

[28] Guy Hoffman. 2019. Evaluating Fluency in Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218. https://doi.org/10.1109/THMS.2019.2904558

[29] Siddarth Jain and Brenna Argall. 2019. Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 1 (2019), 1–23.

[30] Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline Reinforcement Learning as One Big Sequence Modeling Problem. In *Advances in Neural Information Processing Systems*, Vol. 34. 1273–1286.

[31] Shervin Javdani, Siddhartha Srinivasa, and Andrew Bagnell. 2015. Shared Autonomy via Hindsight Optimization. In *Robotics: Science and Systems*. NIH Public Access. https://doi.org/10.15607/rss.2015.xi.032

[32] Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. 2015. The Dependence of Effective Planning Horizon on Model Accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 1181–1189.

[33] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. MOReL: Model-Based Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 21810–21823.

[34])]kumar2019stabilizing Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. [n. d.]. Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. In *Advances in Neural Information Processing Systems* (2019), Vol. 33. 11784–11794.

[35])]kumar2020conservative Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. [n. d.]. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems* (2020), Vol. 34. 20132–20145.

[36] Hanna Kurniawati, David Hsu, and Wee Sun Lee. 2008. SARSOP: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces. In *Robotics: Science and Systems*, Vol. 2008.

[37] Tor Lattimore and Marcus Hutter. 2014. General time consistent discounting. *Theoretical Computer Science* 519 (2014), 140–154.

[38] 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.

[39] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. (2020). arXiv:2005.01643.

[40] Elita A. Lobo, Mohammad Ghavamzadeh, and Marek Petrik. 2021. Soft-Robust Algorithms for Batch Reinforcement Learning. (2021). arXiv:2011.14495.

[41] David Lopez-Paz, Philipp Hennig, and B. Schölkopf. 2013. The Randomized Dependence Coefficient. In *Advances in Neural Information Processing Systems*, Vol. 26.

[42] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[43] Omid Madani, Steve Hanks, and Anne Condon. 1999. On the Undecidability of Probabilistic Planning and Infinite-Horizon Partially Observable Markov Decision Problems. In *AAAI/IAAI*. 541–548.

[44] Mausam and Andrey Kolobov. 2012. Planning with Markov Decision Processes: An AI Perspective. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–210.

[45] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[46] Stefanos Nikolaidis, Swaprava Nath, Ariel D Procaccia, and Siddhartha Srinivasa. 2017. Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 323–331.

[47] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. 2015. Efficient Model Learning from Joint-Action Demonstrations for Human-Robot Collaborative Tasks. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 189–196.

[48] Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. 2017. Human-Robot Mutual Adaptation in Shared Autonomy. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 294–302.

[49] Marek Petrik and Reazul Hasan Russel. 2019. Beyond Confidence Regions: Tight Bayesian Ambiguity Sets for Robust MDPs. In *Advances in Neural Information Processing Systems*, Vol. 32.

[50] Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. 2023. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[51] R. Tyrrell Rockafellar and Stanislav Uryasev. 2002. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance* 26, 7 (2002), 1443–1471. https://doi.org/10.1016/s0378-4266(02)00271-6

[52] Eric Rosen, David Whitney, Michael Fishman, Daniel Ullman, and Stefanie Tellex. 2020. Mixed reality as a bidirectional communication interface for human-robot interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11431–11438.

[53] Raphaëlle N Roy, Nicolas Drougard, Thibault Gateau, Frédéric Dehais, and Caroline PC Chanel. 2020. How can physiological computing benefit human-robot interaction? *Robotics* 9, 4 (2020), 100.

[54] Malte Schilling, Wolfram Burgard, Katharina Muelling, Britta Wrede, and Helge Ritter. 2019. Shared Autonomy—Learning of Joint Action and Human-Robot Collaboration. *Frontiers in Neurorobotics* 13 (2019), 16.

[55] Gaganpreet Singh, Raphaëlle N. Roy, and Caroline P. C. Chanel. 2022. POMDP-Based Adaptive Interaction Through Physiological Computing. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 32–45. https://doi.org/10.3233/faia220188

[56] Tarek Taha, Jaime Valls Miró, and Gamini Dissanayake. 2011. A POMDP framework for modelling human interaction with assistive robots. In *2011 IEEE International Conference on Robotics and Automation*. IEEE, 544–549.

[57] Kerong Wang, Hanye Zhao, Xufang Luo, Kan Ren, Weinan Zhang, and Dongsheng Li. 2022. Bootstrapped transformer for offline reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 34748–34761.

[58] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. 2020. MOPO: Model-based Offline Policy Optimization. In *Advances in Neural Information Processing Systems*, Vol. 33. 14129–14142.

[59] Wenqian Zhang and Haiyan Wang. 2022. Diagnostic Policies Optimization for Chronic Diseases Based on POMDP Model. In *Healthcare*, Vol. 10. MDPI, 283.