Anticipatory Bayesian Policy Selection for Online Adaptation of Collaborative Robots to Unknown Human Types

O. Can Görür
DAI-Labor
Technische Universität Berlin
Berlin, Germany
orhan-can.goeruer@dai-labor.de

Benjamin Rosman
CSIR, and University of the Witwatersrand
Johannesburg, South Africa
brosman@csir.co.za

Sahin Albayrak
DAI-Labor
Technische Universität Berlin
Berlin, Germany
sahin.albayrak@dai-labor.de

ABSTRACT

As a key component of collaborative robots (cobots) working with humans, existing decision-making approaches try to model the uncertainty in human behaviors as latent variables. However, as more possible contingencies are covered by such intention-aware models, they face slow convergence times and less accurate responses. For this purpose, we present a novel anticipatory policy selection mechanism built on existing intention-aware models, where a robot is required to choose from an existing set of policies based on an estimate of the human. Each of these intention-aware robot models anticipates and adapts to a different human's short-term changing behaviors. Our contribution is the Anticipatory Bayesian Policy Selection (ABPS) mechanism which selects from a library of different response policies that are generated from such models, and converges to a reliable policy after as few interactions as possible when faced with unknown humans. The selection is based on the estimation of the human in terms of long-term workplace characteristics that we call types, such as level of expertise, stamina, attention and collaborativeness. Our results show that incorporating this policy selection mechanism contributes positively to the efficiency and naturalness of the collaboration, when compared to the best intention-aware model in hindsight running alone.

KEYWORDS

Human-Robot Collaboration, Anticipatory Policy Selection, Human Type and Intent Inference, Socially Collaborative Robots

ACM Reference Format:
O. Can Görür, Benjamin Rosman, and Sahin Albayrak. 2019. Anticipatory Bayesian Policy Selection for Online Adaptation of Collaborative Robots to Unknown Human Types. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

Recent advancements in robotics are enabling more human-robot teams to work together for increased productivity. For this purpose, research into human-robot collaboration (HRC) has been mainly inspired by human-human teamwork, the core of which lies in an ability to adapt one's behaviors to the other collaborators by categorizing their observed behaviors. By doing so, humans select appropriate behavioral responses to maintain a reliable and efficient collaboration [13]. Our motivation here is to implement a similar mechanism for robots to ensure their autonomous adaptation to different humans having naturally changing intentions, preferences and behaviors. We call such robots *social cobots* [10].

Towards building such robots, many approaches have been proposed, most of which model human intentions and behaviors as a latent variable in robot planning [3–6, 8–10]. An important open problem of such intention-aware models, for their usability in real life scenarios, is the degree to which they allow for interacting with different humans that change their intentions (goals) [1]. A limitation for such models is that they become computationally expensive and less accurate as a wider variety of human behaviors are modeled [12, 22]. Therefore, for more efficient decision-making, existing models implicitly make the assumptions that a human's intention and collaboration preferences are constant or always relevant to an assigned task [10]. This majorly limits a human's intention and behavior space, whereas in reality a human's dynamic desires and emotions introduce greater uncertainty in human behaviors over the course of repeated interactions [19]. Failing to adapt would restrict the fluency of the collaboration also leading to distrust and frustration from the collaborating human [14].

Our belief is that it is very difficult to design and/or learn a single model for a person a robot is collaborating with, let alone for different human types. A robot could face various type of human behaviors. A human behavior may change as a reaction to robot responses. For example, a human may become less collaborative when a robot frustrates her by interfering with a task that she would not trust the robot with. Such human behaviors may also adapt to the context, such as a different task to collaborate on, changing working conditions, daily mood, etc. [7]. Those changes may even be as a result of another human worker starting to collaborate (e.g. a work shift). To ensure long-term usability of robots, a robot should adapt to both short-term changes in a human collaborator's mental state (e.g. tough day at work) and long-term personal habits, preferences and trust. We call each different combination of such long-term behaviors a unique *human type*. Our intuition is that rather than a single adaptive model, sometimes a robot may need to follow completely different decision-making strategies, i.e. policies, to enable fast and reliable online adaptation to various human types.

In this paper, we present a novel anticipatory policy selection mechanism built on top of existing intention- and situation-aware models for an extended adaptation of robots to various human types. In our previous study, we designed a partially observable Markov decision process (POMDP) that adapts to a human's short-term changing behaviors, modeling her availability, intention (motivation) and capability as a latent variable [10]. Our focus in this study is on a robot's adaptation to human long-term behaviors, i.e, human

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

types. We create a policy library by randomly constructing different robot models based on our existing model design. Through this random generation, we are agnostic to specific human types and behaviors modeled. Our contribution is an Anticipatory Bayesian Policy Selection (ABPS) mechanism based on Bayesian Policy Reuse (BPR) [25], which selects a policy from the library in short time and converges to a reliable and nearly optimal policy after as few interactions as possible. The selection is based on a human's estimated long-term workplace characteristics, such as level of expertise, stamina (or fatigue), attention and collaborativeness[1], that correlate to the policy performance. Instead of modeling known human types as a latent variable, we estimate unknown human types from the observed human behaviors using Bayesian belief estimation. To our knowledge, this is the first time such a policy selection mechanism has been proposed complementing intention-aware planning approaches in HRC, providing fast and reliable anticipatory decision-making for both long-term and short-term adaptation to unknown human types (through ABPS) and their changing behaviors through each selected policy.

Our goal is to show that integrating such a policy selection mechanism contributes positively to the efficiency (e.g. time to finish a task, success rate) and naturalness (e.g. a human's increased willingness to collaborate) of the collaboration, when compared to the best intention-aware model in hindsight running alone. We consider a simulated HRC scenario at a conveyor belt for the task of inspecting and storing various products, each of which has different weights. Different types of modeled humans, responsive to both robot actions and changing environment, collaborate on the task autonomously with our adaptive robot decision-making mechanism implementing ABPS (Section 3). We present our experiments and analysis on our policy selection through its effects on the efficiency and the naturalness of the collaboration (Section 4).

2 RELATED WORK

Human-robot interaction studies have lately focused on human intention-aware robot decision-making models for anticipatory adaptation of the robots to humans. Most of these models introduce human intentions as a latent variable in a POMDP, which causes great complexity with an increasing number of human intentions anticipated and handled. For a reasonable convergence time, such a design conventionally has to limit the human intention space and systemic errors a human can make [12]. Therefore, the studies implicitly make the assumption that either a human's intention (or goal) is constant or it is changing in a known limited intention space [10]. We were unable to find any studies that consider the human behaviors as freely stochastic in an unconstrained environment. As a result, robot decision policies generated by such complex models have been developed and tested under constrained environments with rather limited interactions [2–6, 15]. It has been stated that such assumptions limit a robot's anticipation of a human's dynamic behaviors and goals that mostly occur in the long-term as a result of changing preferences, habits, needs and trust [1, 10, 19].

There has been a handful of studies that removes such assumptions on human intentions and behaviors. Contingencies in human actions have been partly considered [11, 18]; however, all actions are still assumed to be toward fulfilling a task, possibly in a way that differs from the expected plan. In our recent work, we conceptualize an anticipatory decision-making model (a POMDP) for the robot that removes those assumptions and handles a human's unexpected behaviors after the human's changing availability, motivation and capability in collaboration tasks [10]. Even though we show that such a proactive model performs better than a robot model with the assumptions, the handled behaviors are limited and less dynamic, and the POMDP model handles only the basic type of behaviors (i.e., tired, distracted, incapable) as a latent variable. In other words, the robot model was not adapting to different humans but instead acting proactively against one simulated person randomly generating such short-term changing behaviors. In this study, we extend a robot's adaptation to anticipate a variety of human types.

Towards incorporating more variety in human characteristics, some studies have proposed complementary solutions to be built on top of a robot's intention-aware planner to foster high-level strategies. In [21], humans are clustered from observations during a training phase into a finite number of human types. The estimated human type is again used as a latent variable in a MOMDP (mixed observability MDP) model to decide on robot actions. The number of types in this study is a limiting factor, where each different type is considered as a partially observable state. This limitation is majorly due to MOMDPs struggling to scale to more states when each type is introduced as a latent state variable. It has been recently stated that when POMDPs are used to optimize spatio-temporal assignments of robots, accurate system models are needed to evaluate both actions and rewards, which are often unavailable or fail to anticipate and adapt to various conditions in long-term missions [20].

To overcome the limitation of a Markov decision process in its larger scale adaptation, the authors in [3] build several such robot models with varying reward and transition functions to handle different tasks. In other words, the robots are given the ability to explore different policies and trade-off toward higher interaction and task quality. However, the study is limited to analyzing different policies to govern such varieties in humans in the context of pedestrian-robot vehicle interaction leaving out the autonomous selection of an optimal one. Our approach brings together the idea of generating many such reliable Markov models [3] to construct a policy library, and the idea of estimating human types on a meta-level as a complementary solution to the intention-aware models [21]; and goes beyond them to offer a fast and reliable policy selection mechanism as part of a closed-loop robot system. Our policy selection replaces the conventional method of using hierarchical or more complicated POMDPs, as it acts as a discretization of the POMDP models instead of modeling types as a latent variable. This allows us to deal with the problem in a more computationally efficient way, and to handle unknown human types while mitigating the need to learn (expensive) response policies on the fly.

For the policy selection in the context of adaptive social agents, some studies have developed decision trees [17] or Bayesian models, e.g. [22], selecting from a limited number of policies. Towards broader adaptations, a recent study proposes a contextual multi-arm bandit (CMAB) approach in an assistant selection mechanism for a robot [20]. Although this approach has proven to be useful in adapting to human capabilities and constraints in a simulation, the

[1] We use this term to indicate the level of a human's will to collaborate that may change due to, for example, task-relevant distrust of the human to the robot.

exploration factor of a CMAB would be very dangerous and frustrating for a human collaborator in real world. In addition, in policy or reward learning algorithms, the learning rate is very difficult to tune and the response time is considerably high for any interaction in real time. Therefore, to satisfy fast learning rates in HRC, which is related to how fast the human behaviors are changing, the studies mostly assume limited human intention space [21, 23]. Particularly, when human workers have their shift changes or when a human drastically exerts different behaviors (e.g. loss of attention, fatigue or injuries in workplaces [7, 16]), learning a new reward function or a policy would take time which is very costly, especially in collaboration scenarios. Moreover, we still need to have an accurate reward and transition model which, in the end, needs to be applicable to all humans being interacted with, and yet again is not realistic to find. In such cases, it is better to reuse a pretrained model rather than spending too much time on training a new one [25].

In an online HRC, a robot's autonomous, reliable and fast response is a direct influence on the fluency and the naturality of the human-robot teaming [14]. Toward more efficient and safe HRC, we believe the Bayesian Policy Reuse (BPR) algorithm is the best fit to our problem [25]. BPR has been shown to perform better than a multi-arm bandit (fast and reliable policy selection) in online adaptation tasks when faced with a greater uncertainty about the description of the task. It considers *a priori* information leading to less exploration and so less unreliable responses of the robot during operation. In our solution, ABPS, we have updated BPR to incorporate anticipation of a human's uncertainty in her long-term behaviors and to be a generic and complementary solution to the existing intention-aware planning solutions for an increased adaptation in real time. Even though ABPS is agnostic to any labels of human types and robot policies, for our domain we generalize some characteristic features of humans in workplaces, inspired from [7, 16, 20], that are crucial for a collaborative robot to know. These are a human's expertise, attention, stamina-level and collaborativeness and they are used to describe a human type.

3 METHODOLOGY

3.1 Overall Framework

Figure 1 shows our overall framework with an overview of how we organize the decision-making process in a human-robot collaboration. In the upper layer, tasks are created and assigned to either the human or the robot. In the meta-cognitive level the *ABPS agent* selects one decision strategy among a *policy library* according to the anticipated type of the human partner. Following the selection, a decision strategy is forwarded to the cognitive level for the *decision-making agent* to execute. In our implementation, each decision strategy is a robot policy comprised of optimal actions for each possible belief over the world states and generated when a POMDP robot model is solved for maximizing expected rewards (see in Section 3.2.1).

The cognitive level of the system in Figure 1 has been the focus of similar studies, which involves a robot's decision-making agent acted upon one precomputed anticipatory model, in our case a POMDP model [10]. In this work, we focus on the *meta-cognitive level*. It includes the policy library constructed from such handcrafted Markov models (detailed in Section 3.2.1) and our

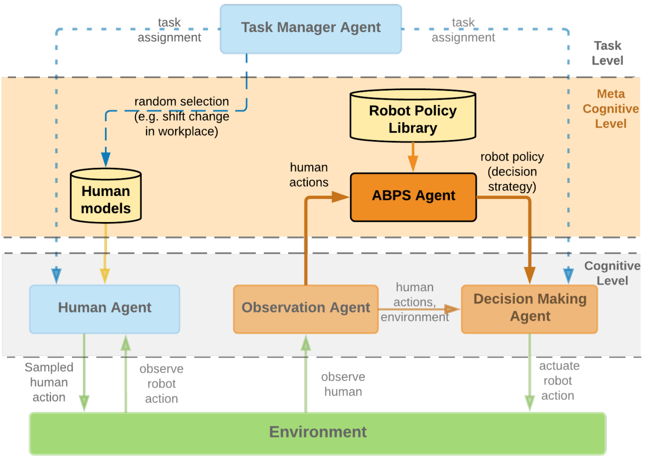

Figure 1: Anticipatory Bayesian Policy Selection (ABPS) agent in the overall framework of our autonomous system

ABPS mechanism consisting of human type (belief) estimation (Section 3.2.2) and policy selection with an exploration heuristic for a quick adaptation to a class of human types (Section 3.2.3).

3.2 Anticipatory Bayesian Policy Selection (ABPS)

Our approach is based on the Bayesian policy reuse algorithm [25]. The definition of ABPS is given with Definition 1.

DEFINITION 1 (**ABPS**). *An ABPS agent is equipped with a policy library* Π *to act appropriately in the context of some human types and tasks in HRC domain. The agent is presented with a human collaborator having an unknown type in a known task, which must be solved within a limited time and small number of trials. The goal of the agent is to select policies from* Π *for the new and possibly unknown human type, over which it has a belief distribution* $\beta(.)$*, while minimizing the total regret in a limited time. Minimizing the regret in this domain is defined by increasing the task success rate and decreasing the amount of warnings received from a human collaborator, relative to the best alternative from* Π *in hindsight.*

ABPS measures the similarity between an unknown human type and previously known types to identify which policies may be the best to reuse. In this case, a collaborated human's type is latent and the human type space is not fully known. Therefore, a correlation between policies and a bounded set of human types is not possible. The similarity of types is extracted from offline training with some known types and by utilizing this trained model online, constructing $\beta(.)$. The general algorithm is given in Algorithm 1. We first detail how a policy and the library Π is constructed in Section 3.2.1. The observation signals and the observation model for the human type belief update (see *line 7, 8* of Algorithm 1) are detailed in Section 3.2.2. Then, the policy selection step and the construction of the performance model used in this step (in *line 4* of Algorithm 1) is described in Section 3.2.3.

3.2.1 Policy Library Construction. To generate many policies for the policy library, we use the anticipatory robot model design simplified in Figure 2 as a base, which we have previously shown

Algorithm 1 Anticipatory Bayesian Policy Selector (ABPS)

```
Require: Human type space $\mathcal{T}$, robot policy library $\Pi$, an observation vector for observed human behaviors $\sigma$ in an observation space $\Omega$, an observation model to match observables to known human types $P(\Omega|\mathcal{T},\Pi)$, utility as accumulated discounted reward obtained from running a policy $U$, a performance model $P(U|\mathcal{T},\Pi)$, number of tasks $K$, exploration heuristics $\mho$.
Train offline for performance and observation models.
Initialize a belief: $\beta^0$ uniform distribution from the prior $\mathcal{T}$.
for task IDs $t = 1 \ldots K$ do
    Select a policy $\pi^t \in \Pi$ using $\beta^{t-1}$ and performance model, $P(U|\mathcal{T},\Pi)$, using $\mho$ in Equation (3).
    Apply selected policy $\pi^t$ to the task and the human.
    wait(until the task $t$ is completed)
    Obtain observations $\sigma^t$ from the human and the environment emitted during the task.
    Update belief $\beta^t$ using $\sigma^t$ by belief update function in Equation (1).
end for
```

to perform well in this context [10]. The model is the POMDP tuple $\{S, A, T, R, \Omega, O, \gamma\}$. S comprises the hidden states of a human's task related availability, motivation and capability. These are anticipated at the first stage (see *Stage-1* in Figure 2). Then, moving from *Stage-1* the robot anticipates whether the human needs assistance or not from the robot's perspective at *Stage-2*. The other states are the global success and failure states that define the result of a task (terminal states), the states of a new task assigned to the agents (initial states), and a state when the robot receives a warning from the human for any reason. A is the robot actions to wait for human (*idle*), plan for assisting action (*planning*) and *assist* human as shown in Figure 2. T is the state transition probabilities. R is the immediate reward the robot receives. Positive rewards are acquired when a task has been accomplished by any agent and negative rewards are for a task failure or when warnings are received from the human. The latter is to encourage the planning to be less intrusive, i.e., the robot will not offer assistance unless it is deemed part of the optimal policy. Ω is the set of human action and task observations as detailed in Section 3.2.2 and O represents the conditional observation probabilities. γ is the discount factor for delayed rewards and we solve the model for an optimal robot policy, π.

The offline generation of different policies to construct the policy library is done by adjusting T and O: the state and observation probabilities of the model corresponding to different human types. Changes in T correspond to different transitions of a human's internal states, e.g. a robot policy assumes the human tires faster (related to the stamina-level) or the human needs assistance when she is not capable (related to the collaborativeness). Whereas changes in O define the observations emitted by the human as a function of her internal states. For example, a human not being able to handle the task could indicate that she is tired, or she is a beginner (related to expertise) depending on her type, both of which should be handled differently by the robot. Additionally, by adjusting O, we are able to make the model a partially observable, mixed observability or fully observable Markov decision process (POMDP, MOMDP or MDP, respectively). We randomly adjust the probabilities as mentioned above to generate various Markov decision models, each of which handles a unique human type, and solve for their optimal policies to construct our policy library Π. The main reason we move from a base model as in Figure 2 is to limit the arbitrary generation of robot policies to avoid overloading the space with unreliable candidates [1]. This way we also show how we integrate ABPS to existing intention-aware models.

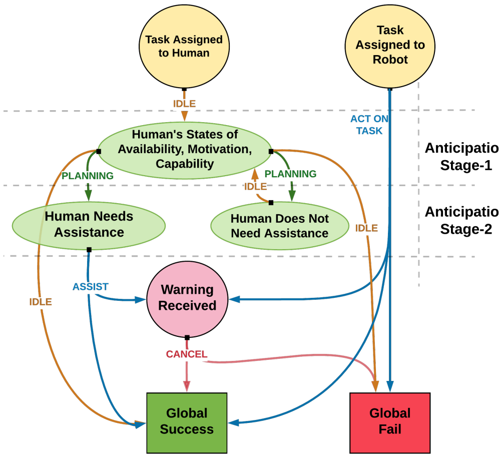

Figure 2: Our anticipatory robot model design as a Markov decision model. At the first stage, the robot anticipates human states, such as *human may be tired, may not be capable, doing fine.* Then, it anticipates whether the human needs help moving from first stage estimations [10].

3.2.2 *Human Type Belief Estimation.*

The space of human types is in general infinite, but we limit this to control complexity. Therefore, the construction of a type space $\mathcal{T}$ is a crucial process. For this purpose, we train an estimation model from a set of known types and use it online to estimate a new unknown type as a belief distribution over the known ones, $\beta(.)$. In order to train such a model, we generalize some characteristic human features to approximate a human type. These features are inspired from [7, 16, 20] and are stated to be crucial to be known by a collaborative robot. These are a human's expertise, attention, stamina-level and collaborativeness. The last term is a more general description of a human's acceptance rate of a robot's offer for assistance. The type space consists of many human types by adjusting the level of these features, e.g. a human with beginner skills, pensive, bad stamina and non-collaborative behaviors (e.g. always rejecting a robot's assistance due to distrust). We argue that any human worker can be represented as a distribution of such features in our experiments. More details on the simulated human types in type space $\mathcal{T}$ are given in Section 4.1.2.

The human type estimation model is used by ABPS as *a priori* information, which we call the *observation model.*

Definition 2 (**Observation model**). *For a robot policy π, a human type τ and an observation vector σ obtained from the human actions and the environment, the observation model $P(\sigma|\tau,\pi)$ is a probability distribution over the observation signals $\sigma \in \Omega$ that results by applying the policy π to the type τ.*

All the combinations of known human types in τ and the robot policies in the library are run against each other offline several times to generate our *observation model* (detailed in Section 4.1.3). The observation signals are emitted by the collaborated human and the environment, reflecting a human's actions and their impact on the task and the environment. In our experiments, an observation vector, $\sigma \in \Omega$, is a 6-D boolean vector with the following observables: $\{$*human is detected, human is looking around, human has taken a task related action and succeeded in it (e.g. grasping and lifting a package in our scenario), human has taken a task related action and failed, human is warning the robot, human is idle*$\}$. The ABPS agent receives these observables at every episode of a task and accumulates them to update its belief on the human type after a task finishes (see *line 6, 7, 8* of Algorithm 1). Finally, the type belief update is Bayesian, given by

$$\beta^t(\tau) = \frac{P(\sigma^t|\tau, \pi^t)\beta^{t-1}(\tau)}{\sum_{\tau' \in \mathcal{T}} P(\sigma^t|\tau', \pi^t)\beta^{t-1}(\tau')}, \quad \forall \tau \in \mathcal{T} \tag{1}$$

where β^{t-1} stands for the previous belief and $P(\sigma^t|\tau, \pi^t)$ is the probability of observing σ^t after applying π^t in an interaction with any human type τ. This distribution is directly retrieved from the *observation model* for each requested type and policy.

3.2.3 Policy Selection with Exploration Heuristics. The policy selection process of the robot is based on an exploration heuristic called *expected improvement (EI)* [25]. As stated in line 4 of Algorithm 1, this algorithm runs on another trained *a priori* model called the *performance model.*

Definition 3 (**Performance model**). *The performance model, $P(U|\tau, \pi)$, is a probability distribution over the utility, U, of a policy π when applied to human type $\tau \in \mathcal{T}$.*

The system utility, U, is the accumulated discounted reward received after a policy is run (see Section 3.2.1 for the immediate rewards a robot obtains during a task). All the combinations of known human types $\tau \in \mathcal{T}$ and the robot policies $\pi \in \Pi$ are repeatedly run against each other offline to generate our *performance model.* Then, this model is used by the policy selection heuristic.

The heuristic assumes that there is a U^+ in reward space which is larger than the best estimate under the current type belief, U^β. A probability improvement algorithm can be defined to choose the policy that maximizes Equation (2) and achieves the utility U^+.

$$\pi' = \arg\max_{\pi \in \Pi} \sum_{\tau \in \mathcal{T}} \beta(\tau) P(U^+|\tau, \pi) \tag{2}$$

Because the choice of U^+ directly affects the performance of the exploration, its selection is crucial to the performance of this exploration. The *expected improvement* approach instead addresses this nontrivial selection of U^+. The algorithm iterates through all the possible improvements on an existing U^β of the current belief, which satisfies $U^\beta < U^+ < U^{max}$. The policy with the best potential is then chosen, as given in Equation (3).

$$\pi' = \arg\max_{\pi \in \Pi} \int_{U^\beta}^{U^{max}} \sum_{\tau \in \mathcal{T}} \beta(\tau) P(U^+|\tau, \pi) dU^+ \tag{3}$$

$$= \arg\max_{\pi \in \Pi} \sum_{\tau \in \mathcal{T}} \beta(\tau)(1 - F(U^\beta|\tau, \pi)) \tag{4}$$

where $F(U^\beta|\tau, \pi) = \int_{-\infty}^{U^\beta} P(u|\tau, \pi) du$ is the cumulative distribution function of U^β for a τ and π. The algorithm, therefore, selects the robot policy with the most likely improvement on the expected utility.

4 EVALUATION

4.1 Experiments

In this section, we first detail the simulation environment we have used for our experiments in Section 4.1.1. We then give our human simulation mechanism and how human models are crafted towards simulating short and long-term changes in human behaviors and types, in a task of inspection and storage of various products (Section 4.1.2). After that, we describe the training phase to construct the library and the estimation models for ABPS (Section 4.1.3). Finally, we give the details of how we conduct the real-time experiments and our performance metrics in Section 4.1.4.

4.1.1 Simulation Environment. We implement the proposed architecture in Figure 1 in the Robot Operating System (ROS) and use our simulation environment developed under the MORSE environment. As seen in Figure 3, we have developed an updated version of the MORSE human and PR2 models with special actions related to our use-case task. The simulation environment allows our robotic system to run a long-term collaboration. Such long-term experiments make it possible for the robots to face many different changing human types and behaviors under various conditions. As a result, we do not have to be limited to constrained environments and human interactions. This helps us train very accurate models of the interaction, as well as run rigorous tests on our system facing and covering more uncertainties of humans.

All of our scenarios consist of several sequential task assignments to simulate a long-term collaboration. A task in our case is product inspection and storing. It starts with an initial task assignment to either robot or the human based on the product's weight and fragility. We only consider the cases where a task is assigned to the human, in order to keep our focus on anticipating the human's type and behaviors and correctly assess her need for assistance. The collaboration is when the robot correctly estimates the human's such need and helps with the task. A task is successful when the product is inspected and put into green containers either by the human or by the robot (see Figure 3a). We set a maximum allowed processing time for each product inspection, t_{max}, to keep the collaboration and production flowing in the factory. The conveyor belt waits for t_{max} for a package to be processed, or else it runs and the product falls into the uninspected-product container (the red container in Figure 3a) leading to a task failure. As stated in Algorithm 1, a new policy is selected after each task is finalized.

4.1.2 Human Simulation. We have modeled many different human types for our collaboration scenarios. In our experiments, we run randomly generated models to reflect changing and unknown levels of expertise, stamina, attention and collaborativeness. The models reflect them as actions, which are observations for the robot obtained from 3D human body joints always available directly from the simulated humans. Since it is not the focus of this study, we use a state-of-the-art human activity recognition (HAR) system inspired by existing studies, e.g., [24], to recognize the constrained and distinctly simulated human gestures: the human is looking around (i.e. distracted as in Figure 3a), attempting the task (see Figure 3d), warning the robot (a special gesture to stop the robot as shown in Figure 3b), idle (inactive), walking away (see Figure 3c). During a task, the robot collects all the observations emitted from

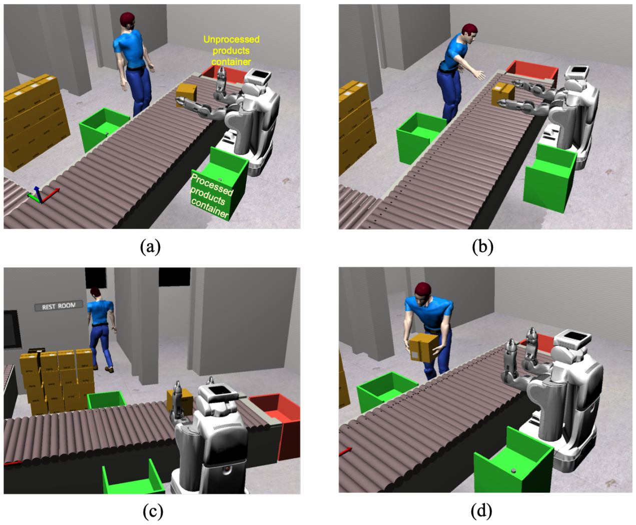

Figure 3: Our HRC scenario. (a) Distracted human while robot is pointing out to remind, with containers shown; (b) Robot takes over the task to assist and human gestures to stop the robot (warns the robot); (c) Human walks away for a rest; (d) Idle robot while human is grasping the product.

the human and the environment and averages them to construct our 6-D observation vector for the task (given in Section 3.2.2). The observation vector also has features of the human's success in the task. After each attempt the human makes at grasping, the robot associates the human's attempts with expertise.

In our experiments, we assume that a human worker optimizes an objective function to reach her goal. However, following our statement, this may also be an internal goal irrelevant to the assigned task, e.g. leaving her place for a short break. We also assume that any human actions towards her goal may be imperfect [12]. Simulating such a human has been shown to be accurate using a Markov decision process (MDP) to generate a policy for a human agent [3, 10]. For this purpose, we use an updated version of our human model in [10] in our simulations, which is inspired by studies on human behaviors in a workplace [7, 16, 20]. Our human MDP is a tuple $\{S, A, T, R, \gamma\}$ where S is the human states of mind, A is the human actions, T is the state transition probabilities and R is the immediate rewards received based on the result of a task and the type of the human to encourage that type of behavior. For example, a beginner and non-collaborative human model receives positive rewards when the human cannot handle the task and each time the human warns the robot when the robot interferes with the task. The model is inspired by our expectations that a human chooses an action based on the collaborated robot's action, the state of a task, the human internal states and human internal goals.

Our update to the human model in [10] is toward governing the human's responsiveness to the interacted robot actions. Such responsiveness is handled through a transition function $T(s, a, s') = P(s'|s, a, n_r, k_t)$ for $s, s' \in S$, $a \in A$, the number of times the robot interfered in a task n_r and the number of tasks handled so far k_t. That means we have dynamic transition probabilities changing over the course of the interactions, leading to updates on human models after each task, and so updated human behaviors. An example to such responsive behaviors is that a human becomes less collaborative as her robot partner selects wrong policies, e.g., the robot takes over a task (depicted as n_r) when the human was already planning to handle it. A decrease in collaborativeness is handled with an increased transition probability of the human to *warn the robot* when a robot interferes with a task. Another example is that a transition to the state of being tired depends on the number of tasks already handled, k_t. Finally, each human model is simulated with random sampling using Markov Chain Monte Carlo (MCMC). In the end, the MCMC simulation and the responsive transition function lead to human simulations exerting dynamic behaviors changing in response to the robot decisions and with a small random factor. Through this modeling scheme we create various human types with changing characteristics, e.g. *a beginner, tired and collaborative human*, a *mid-expert, high-stamina and less-collaborative human.*

4.1.3 *Training Phase.*

We have generated many different robot policies to build Π, the library. During the generation, all robot model designs have random transition and observation probabilities assigned (see Section 3.2.1); therefore, they are agnostic to the state transitions inherent in the human models. In the end, 20 policies have been selected for their use in the experiments in order not to overload the policy library. This selection is based on how well they performed overall against many different randomly generated human models, i.e., discarding the worst ones, and how distinct their *performance models* are (see Definition 3) from the other policies, i.e., grouping the similar ones. Some policies ignore a human's warnings and try to complete a task, whereas some pay more attention to a human's needs, taking the human as the leader of the collaboration. The trade-off between these two is more obvious when it comes to non-collaborative types. There are also some policies that prefer to encourage the human to complete the task, e.g. by pointing out to remind the human when distracted instead of directly taking over the task. Which policy is more optimal depends on the interacted human type and the task definition.

We are agnostic to the exact type labels of humans in our experiments. As mentioned, we assume each human the robot is interacting with has an unknown type to the robot, which can only be estimated as a distribution over the known types. For this purpose, we have crafted 16 different human types (known types), again with the goal of each generating as distinct set of observations (human actions) as possible toward a more heterogeneous distribution. Generating distinct observations means creating human types with extremes of the four features, namely the levels of expertise, stamina, attention and collaborativeness. Our assumption is that an unknown human type can be approximated as a probability distribution over these extreme types. Increasing the number of known human types would yield less accurate type estimations with a higher convergence time. However, we note that since each of the 16 human models are stochastic, they still generate a diversity of behaviors after random sampling (see Section 4.1.2).

During the training phase we run each of the 20 robot policy against the 16 known human types for 50 sequential tasks. In total, human and robot models accomplished 16000 interactions (16000 task instances), which is very difficult to manage in real-life scenarios. The *performance model* and the *observation model* (see Definition 2) are constructed after this training phase.

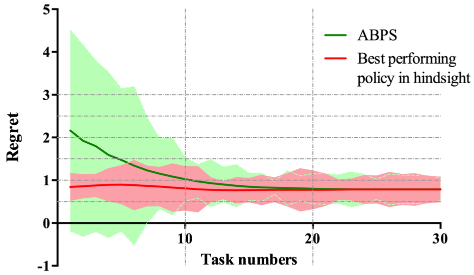

Figure 4: Moving average regret over time with error bars denoting the standard deviation, collected by ABPS and the best performing policy in hindsight for the experimented human type.

4.1.4 Real-Time Experiments Phase. Our goal is to prove the hypothesis below:

A single intention-aware robot model is limited in its adaptation to various human types. Our ABPS mechanism provides broader adaptation to various and changing human types leading to more efficient and natural collaboration, while maintaining fast and reliable convergence to the best policy.

To explore the hypothesis, we conduct two lines of experiments and gather the objective measures below:

Experiment-1: *The goal of this experiment is to see the performance of ABPS in terms of how fast and reliably it converges to the best policy performance.* For this purpose, we compare an ABPS robot collaborating with a randomly created human type unknown to the robot, with the best performing robot policy for that type when collaborating with the same human. The best robot policy is picked from the library as the best performer in hindsight for the experimented human type. Both robots interact with the human for 30 sequential task assignments and this scenario is repeated 10 times for each. We measure the moving average regret of both of the robots and compare the change over time. A regret for a selected policy $\pi \in \Pi$ is $R^{\tau}_{\pi} = U^{\tau}_{\pi *} - U^{\tau}_{\pi}$ for a human type $\tau \in \mathcal{T}$ and the best policy $\pi * = \arg\max_{\pi \in \Pi} U^{\tau}_{\pi *}$.

Experiment-2 *The goal of this experiment is to show the contribution of our ABPS model to the adaptation capability of a robot through increased efficiency and naturalness in the collaboration.* For this purpose, we compare the performance of ABPS robot with the library's best policy in hindsight running alone, when collaborating with various types of humans unknown to the robot and changing during the operation. The robot is unaware of this change, which might be thought of as a shift change in a factory environment. The best policy for this experiment is the overall best performer in the policy library, picked after the training phase when averaged over all the interactions. For the purpose of simulating unknown human characteristics, we randomly crafted 10 different human types offline. At every 30th task in a scenario (enough to let ABPS converge), the human type changes drastically to another unknown human type in a certain order, and the robot has to adjust its responses accordingly. The same order is repeated 5 times (300 sequential tasks in each scenario) for each strategy to average and smooth the human type characteristics and observe the long-term performance of the robots. We analyze the following for both of the strategies:

- *How the human state distribution and the average reward the robot collects change over time.* This shows the effect of such type and behavior changes on a single intention-aware model that introduces those changes as a latent variable versus the ABPS mechanism and its adaptation capability.
- *Success rate, the number of warnings the robot received from humans and the approximate time a task takes.* These are to compare the task efficiency and naturalness of the robots. We also analyze the trade-off between time and success rate as made by the policy selector to avoid human warnings.

4.2 Results

The results of *Experiment-1* are illustrated in Figure 4. ABPS naturally has a uniform belief distribution over the human types when first initialized, and has selected different policies (best performers of the library) until its belief estimation converges (as shown by higher deviations). It has already reached a very close performance to the best policy after the 6th task, by correctly selecting the same policy at that time (i.e., at the 6th iteration). The difference between the moving average regrets of both strategies decreases to equalize after this point showing the fast convergence of ABPS (one-way ANOVA: $F(1, 58) = 0.017, p = 0.895$). It should be noted that a zero regret cannot be reached even by the best policy and there is a constant variance on the values. This reflects that the human's behaviors are constantly changing over the course of the interaction. In a real world setup we may have more stable behaviors from people; however, with this experiment we point out the adaptation performance of ABPS.

For *Experiment-2*, the results are shown in Figure 5 and Figure 6. We compare our ABPS with the overall best policy in hindsight in the policy library. To reflect the dynamic nature of our human simulations, Figure 5a shows the average duration of each different human state in one task, and how this duration changes over the task assignments. Within every 30 tasks we visualize a human's changing availability, motivation and capability which are generated by a single type. These behaviors are reflected under the human states of *failed to handle* the task, being *tired*, being *distracted*, *evaluating* (spending time to figure out how to achieve) and *warning the robot* (when a human does not want the robot's assistance) as shown in Figure 5a. The drastic changes of these states after every 30 tasks shows the different long-term characteristics, i.e., types, of human workers. For example, the human which took a shift between the 90-120th tasks is more of a distracted type whereas between the 210-240th tasks is a more expert human, with less failures and evaluating time.

We note again the dynamic nature of each human models and Monte Carlo sampling yield various behaviors as shown in Figure 5a, e.g. an expert human can also fail sometimes. This causes many fluctuations in the moving average rewards collected after each task as illustrated in Figure 5b. It is noticeable how the rewards are affected by different humans starting to interact with the robot (at every 30th task instance). In most of such cases, the overall best policy model is affected more negatively than ABPS. For example, between the 180-210th tasks the human type resembles more beginner and less collaborative behaviors than the others due to the number of warnings she made to the robot and the number of

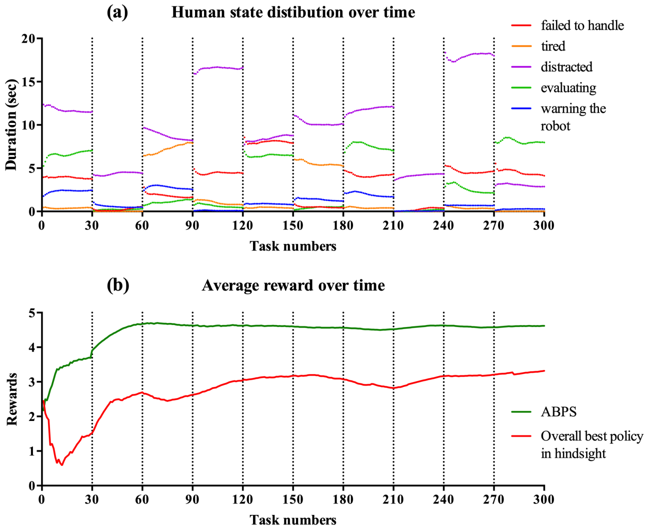

Figure 5: (a) Human state distribution over time showing the changes on human types and behaviors. (b) Moving average reward over time collected by ABPS and the overall best single policy in hindsight (best performer in the library).

failures. Such a difficult human type causes a drop in the average reward the overall best policy collects, whereas our ABPS selects another policy (the best for that type) adapting to the situation that results in almost no change to the reward level. This policy has avoided possible warnings by not taking over the human's task directly but encouraging the human and waiting patiently for the human to handle it unless it is too late for the task. ABPS shows its necessity in such a realistic system, especially at the beginning of the experiment where the overall best policy is clearly not suitable for that type of human. In general, one-way ANOVA tests show that the accumulated rewards of ABPS is significantly larger (see in Table 1) with a mean difference of 39.2%. This and the almost stable reward level in Figure 5b shows that ABPS provides faster and more reliable adaptation to these difficult cases.

As shown in Figure 6b and Figure 5b, the warnings received by the overall best policy accumulate greatly, especially against the humans between the 60-90th and 180-210th tasks (the former is tired, the latter is a beginner and both are non-collaborative), whereas the ABPS robot has successfully adapted to the situations and accumulated fewer warnings. This shows such an adaptation is necessary for the naturalness of the collaboration, which also affects the success rate of the system and finally the accumulated reward being the combination of both (Figure 6c and Figure 6a, respectively). During these task intervals, ABPS trades off the duration of the tasks with the efficiency and naturalness through the selected policies (see Figure 6d and Figure 6b). The human type is likely to exert slightly more non-collaborative behaviors and the policy selection of the ABPS favors avoiding interference, waiting for the human to succeed, collecting more rewards through higher success rates and *fewer warnings*. On the other hand, the best policy offered assistance and took over the tasks in general. This resulted in more warnings received by the best policy and lower success rates as the robot had to cancel its action after the warning. However, it led to faster times of completing the task, most of which were a failure. Despite such small trade-offs, ANOVA tests show that ABPS leads to significantly better efficiency (with 9.5% higher mean success rate and with 14.6% less mean task duration) and more natural (with 58.8% fewer warnings received) human-robot collaboration, as summarized in Table 1.

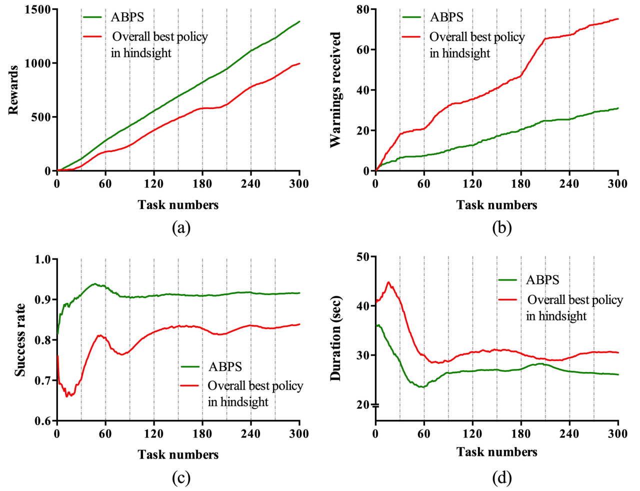

Figure 6: Comparison of the robot with ABPS and the robot with the overall best single policy in hindsight over the task assignments: (a) cumulative rewards acquired; (b) number of warnings received; (c) moving average of the success rate; (d) moving average of the task durations in seconds.

Table 1: Final results from Experiment-2: ABPS vs. overall best policy (μ = *mean*)

Type	μ_{ABPS}	$\mu_{BestPolicy}$	ANOVA
Discounted rewards	4.62	3.32	$F(1, 598) = 74.11$, $p < 0.0001$
Total warnings	31	75.2	$F(1, 598) = 70.37$, $p < 0.0001$
Success rate	0.92	0.84	$F(1, 598) = 29.94$, $p < 0.0001$
Task duration	26.04 $secs$	30.49 $secs$	$F(1, 598) = 15.61$, $p < 0.0001$

5 CONCLUSION

We introduce our novel anticipatory Bayesian policy selection (ABPS), in an HRC setup as a complementary solution to the existing intention-aware robot decision-making models. We examine the effects of our ABPS on a collaborative robot's adaptation to unknown human types and their changing behaviors in a long-term collaboration. Our results have shown that ABPS is a fast and reliable policy selection mechanism for HRC scenarios. Having such a mechanism on top of a robot's intention-aware decision-making contributes positively to the efficiency and naturalness of the collaboration by providing better adaptation to the collaborated human, when compared to the state-of-the-art robot decision-making models running alone. In our experiments, we have utilized our fully autonomous architecture capable of running ABPS along with a robot's decision-making and observation agents, in a real-time human-in-the-loop simulation setup. In future work we will validate our system through user studies on a real setup.

REFERENCES

[1] Stefano V. Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66 – 95.

[2] Chris L. Baker and Joshua B. Tenenbaum. 2014. Modeling Human Plan Recognition using Bayesian Theory of Mind. In *Plan, Activity, and Intent Recognition: Theory and Practice*. 177–204.

[3] Tirthankar Bandyopadhyay, Kok Sung Won, Emilio Frazzoli, David Hsu, Wee Sun Lee, and Daniela Rus. 2013. Intention-Aware Motion Planning. In *Algorithmic Foundations of Robotics X*, Emilio Frazzoli, Tomas Lozano-Perez, Nicholas Roy, and Daniela Rus (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 475–491.

[4] Frank Broz, Illah Nourbakhsh, and Reid Simmons. 2013. Planning for Human-Robot Interaction in Socially Situated Tasks: The Impact of Representing Time and Intention. *International Journal of Social Robotics* 5, 2 (2013), 193–214.

[5] 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 (HRI '18)*. ACM, New York, NY, USA, 307–315.

[6] Sandra Devin and Rachid Alami. 2016. An Implemented Theory of Mind to Improve Human-Robot Shared Plans Execution. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. 319–326.

[7] Matthew Gombolay, Anna Bair, Cindy Huang, and Julie Shah. 2017. Computational Design of Mixed-initiative Human-robot Teaming That Considers Human Factors. *Int. J. Rob. Res.* 36, 5-7 (June 2017), 597–617.

[8] O. Can Görür and Aydan M. Erkmen. 2015. Intention and Body-Mood Engineering via Proactive Robot Moves in HRI. In *Handbook of Research on Synthesizing Human Emotion in Intelligent Systems and Robotics*, Jordi Vallverdú (Ed.). IGI Global, 256–284.

[9] O. Can Görür, Benjamin Rosman, Guy Hoffman, and Şahin Albayrak. 2017. Toward Integrating Theory of Mind into Adaptive Decision-Making of Social Robots to Understand Human Intention. In *Workshop on Intentions in HRI at ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*.

[10] O. Can Görür, Benjamin Rosman, Fikret Sivrikaya, and Sahin Albayrak. 2018. Social Cobots: Anticipatory Decision-Making for Collaborative Robots Incorporating Unexpected Human Behaviors. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. 398–406.

[11] Laura M. Hiatt, Anthony M. Harrison, and J. Gregory Trafton. 2011. Accommodating human variability in human-robot teams through theory of mind. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*. 2066–2071.

[12] Laura M Hiatt, Cody Narber, Esube Bekele, Sangeet S Khemlani, and J Gregory Trafton. 2017. Human modeling for human-robot collaboration. *The International Journal of Robotics Research* 36, 5-7 (2017), 580–596.

[13] Guy Hoffman and Cynthia Breazeal. 2004. Collaboration in Human-Robot Teams. In *AIAA 1st Intelligent Systems Technical Conference*. Chicago, IL, USA.

[14] Guy Hoffman and Cynthia Breazeal. 2007. Effects of Anticipatory Action on Human-robot Teamwork Efficiency, Fluency, and Perception of Team. In *Proceedings of the 2007 ACM/IEEE International Conference on Human-robot Interaction (HRI '07)*.

[15] Chien Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. 83–90.

[16] Qiang Ji, Peilin Lan, and Carl Looney. 2006. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 36 (2006), 862–875.

[17] Ece Kamar, Ya'akov Gal, and Barbara J. Grosz. 2009. Incorporating Helpful Behavior into Collaborative Planning. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2 (AAMAS '09)*. 875–882.

[18] Hema S. Koppula, Ashesh Jain, and Ashutosh Saxena. 2016. *Anticipatory Planning for Human-Robot Teams*. Springer International Publishing, Cham, 453–470.

[19] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics* 5, 2 (2013), 291–308.

[20] S. McGuire, P. M. Furlong, C. Heckman, S. Julier, D. Szafir, and N. Ahmed. 2018. Failure is Not an Option: Policy Learning for Adaptive Recovery in Space Operations. *IEEE Robotics and Automation Letters* 3, 3 (2018), 1639–1646.

[21] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. 2015. Efficient Model Learning from Joint-Action Demonstrations for Human-Robot Collaborative Tasks. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. 189–196.

[22] Jan Pöppel and Stefan Kopp. 2018. Satisficing Models of Bayesian Theory of Mind for Explaining Behavior of Differently Uncertain Agents. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. 470–478.

[23] Ramya Ramakrishnan, Chongjie Zhang, and Julie Shah. 2017. Perturbation Training for Human-robot Teams. *Journal of Artificial Intelligence Research* 59, 1 (May 2017), 495–541.

[24] Alina Roitberg, Alexander Perzylo, Nikhil Somani, Manuel Giuliani, Markus Rickert, and Alois Knoll. 2014. Human activity recognition in the context of industrial human-robot interaction. In *2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014*.

[25] Benjamin Rosman, Majd Hawasly, and Subramanian Ramamoorthy. 2016. Bayesian policy reuse. *Machine Learning* 104, 1 (2016), 99–127.