Legible Action Selection in Human-Robot Collaboration

Huaijiang Zhu[1], Volker Gabler[1] and Dirk Wollherr[1]

Abstract— **Humans are error-prone in the presence of multiple similar tasks. While Human-Robot Collaboration (HRC) brings the advantage of combining the superiority of both humans and robots in their respective talents, it also requires the robot to communicate the task goal clearly to the human collaborator. We formalize such problems in interactive assembly tasks with hidden goal Markov decision processes (HGMDPs) to enable the symbiosis of human intention recognition and robot intention expression. In order to avoid the prohibitive computational requirements, we provide a myopic heuristic along with a feature-based state abstraction method for assembly tasks to approximate the solution of the resulting HGMDP. A user study with human subjects in round-based LEGO assembly tasks shows that our algorithm improves HRC and helps the human collaborators when the task goal is unclear to them.**

I. INTRODUCTION

Manufacturers often find it uneconomic to customize assembly robots for tasks with only minor differences. In such scenarios, it is ideal to have humans working side by side with robots and carry out the distinct part of the task that cannot be done by the robot alone. However, such collaboration is susceptible to the ambiguity of the tasks and the imperfect memory of humans. Thus, the robot needs to make its intention clear to the human, ideally without verbal communication, as installing communication modules for the robots can as well be uneconomic.

For example, consider an assembly robot that is limited to a set of nonverbal actions, how should it behave to "tell" the human collaborator which task to carry out? More specifically, given a partially accomplished task and the observed actions of the human collaborator, how can the robot make its next actions *intent-expressive*, or *legible*? To answer this question, we need to first understand how human beings interpret actions of other agents.

Research in psychology suggests that human beings tend to interpret actions as goal-directed [1], [2], i.e. humans attribute goals to other agents, including robots [3], as the causes of their actions. One assumption of action understanding, known as teleological reasoning [1], is based on the principle of rational action [4], which states that actions function to realize goal-states by the most efficient means available. This suggests a formulation of action understanding as inverse planning or inverse reinforcement learning (IRL) [5], [6], [7], where efficiency is defined as maximizing the reward or minimizing the cost the agent receives in the environment. Taking a probabilistic perspective, Baker *et al.* proposed in [8] a framework based on Markov decision process (MDP) for action understanding and use Bayesian inference to compute the posterior probability of a goal, conditioned on observed actions and the environment.

Based on these research results, legibility as a property of actions can be characterized. Dragan *et al.* [9], [10] define a legible motion as one that enables an observer to quickly and confidently infer the correct goal. They point out that while legibility and predictability sometimes can be correlated, they are not the same. A predictable motion is formalized as motion that matches the human collaborator's expectation given a goal. That is, it is efficient with respect to the given cost or reward function for the goal, but a legible motion can be and is usually inefficient. Stulp *et al.* [11] show that legible motions can also be generated using Policy Improvement through Black-Box optimization (PIBBO) [12], a model-free reinforcement learning approach, without knowing the underlying cost functions. They improve the robot's motion through direct trial-and-error interactions with humans to decrease the time the humans need to infer the correct goal.

In this paper, we extend the notion of legibility to multi-step human-robot cooperative assembly tasks where the assembly process is viewed as a sequential decision-making problem similar to the ones studied in [13], [14]. The robot is required to establish a legible policy—a mapping from system states to actions—such that the human collaborator can infer the unknown task goal correctly from the partially built object as early as possible without verbal communication. We will refer to this as the *nonverbal legible assembly problem* in later discussion.

In contrast to motion planning, the "trajectory" of assembly tasks is the building process of the object, which is modeled in a discrete state space and affected not only by the robot but also by the human collaborator. Therefore, it is necessary for the robot to infer the human collaborator's expectation of the task goal and adjust its policy accordingly. As legible actions can be inefficient, we argue that employing legible policies only when the human collaborator has a wrong expectation of the task goal, can avoid unnecessary inefficiency. Moreover, inference of the human collaborator's expectation of the task goal is beneficial especially in scenarios where multiple goals are present, as disambiguating multiple goals simultaneously can be hard. A more practical strategy is to compute the probability distribution over the human collaborator's expectation of the task goal and then choose the legible policy such that only the wrong goal expectation with the highest probability is deviated from.

Our contribution in this work is to unify human intention recognition and robot intention expression in one framework by modeling the nonverbal legible assembly problem as a

[1]All authors are with the Chair of Automatic Control Engineering, Technical University of Munich, Theresienstr. 90, 80333 München, Germany. {h.zhu, v.gabler, dw}@tum.de

hidden goal Markov decision process (HGMDP) [15], a special class of partially observable Markov decision processes (POMDPs), where the goal is the only partially observable state variable. On the basis of the underlying task-related cost, or reward, we construct a special form of reward function that promotes legibility, drawing analogy from previous work of Dragan *et al.*. The robot then maximizes the total reward of legibility it collects during the assembly process.

As solving a finite-horizon HGMDP is PSPACE-complete even for deterministic dynamics [15], another contribution of this work is to propose a myopic heuristic: we first learn legible policies offline in reduced fully observable MDPs, and then estimate the current human collaborator's expectation of the goal online through belief updates in the original HGMDP and adjust the robot's policy accordingly. In addition, we introduce a systematic way of state abstraction for assembly tasks to further limit the size of the state space.

In the remainder of this paper, we first illustrate in more detail the proposed framework in Section II and the state abstraction method in Section III. Then, we describe the human subject experiment and analyze the results in Section IV. Finally, we conclude this paper in Section V.

II. NONVERBAL LEGIBLE ASSEMBLY PROBLEM

We consider a nonverbal legible assembly problem in which the "robot", **R**, has full knowledge of the task goal, while the "human", **H**, does not. Moreover, **R** does not observe **H**'s expectation of the goal directly; rather, it only knows a set of possible goals of **H** and has to infer it from **H**'s actions during the assembly process. **R** maintains a probability distribution over the possible goal expectations of **H** and exploits this information to make the actual goal clear to **H** through its actions without verbal communication.

A. Model Overview

Formally, we model the problem as a HGMDP. Using a factored representation similar to [16], we define it as a tuple $M = (\mathcal{X}, \mathcal{Y}, I_Y, \mathcal{A}^{\mathrm{R}}, \mathcal{A}^{\mathrm{H}}, \mathcal{O}, T_X, T_Y, Z, R^{\mathrm{R}}, R^{\mathrm{H}}, R_{\mathrm{L}}, \gamma, y^*)$.

$\mathcal{X}$ is a finite set of fully observable task states $x \in \mathcal{X}$; $\mathcal{Y}$ is a finite set of partially observable states $y \in \mathcal{Y}$ representing the goal expectation of **H**, whose prior distribution $I_Y(y) = P(y)$ is given; $y^* \in \mathcal{Y}$ denotes the actual task goal which is known beforehand only to **R**. $\mathcal{A}^{\mathrm{R}}$ is a set of actions for **R** and $\mathcal{A}^{\mathrm{H}}$ is a set of actions for **H** that can be observed by **R**, i.e. the set of observations $\mathcal{O} = \mathcal{A}^{\mathrm{H}}$. $T_X(x, y, a^{\mathrm{R}}, x') = P(x'|x, y, a^{\mathrm{R}})$ and $T_Y(x, y, a^{\mathrm{R}}, x', y') = P(y'|x, y, a^{\mathrm{R}}, x')$ are factored transition probability functions of the system.

A transition in this HGMDP proceeds as follows: given a system state $(x, y) \in \mathcal{X} \times \mathcal{Y}$, **R** makes an action $a^{\mathrm{R}} \in \mathcal{A}^{\mathrm{R}}$, resulting in an intermediate task state $\tilde{x}$. Then **H** makes an action $a^{\mathrm{H}} \in \mathcal{A}^{\mathrm{H}}$ according to a stochastic policy $\pi^{\mathrm{H}}(\tilde{x}, y, a^{\mathrm{H}}) = P(a^{\mathrm{H}}|\tilde{x}, y) \mapsto [0, 1]$ and this leads to the next state (x', y'). We assume that the transition of task states is deterministic with respect to the actions a^{R} and a^{H}; the uncertainty arises rather from **H**'s policy.

In modeling the system, we only look at the states where **R** needs to make a decision; the intermediate states and the effect of **H**'s actions are implicitly modeled in the transition probabilities of the system; hence **H** is modeled as part of the environment. For simplicity of notation, it is assumed that the task state x' also encodes the preceding human action a^{H}.

Now, **R** is required to maximize a special form of reward R_{L} promoting legibility in the dynamics defined above. The total reward is discounted in time by the factor γ to give less weight to rewards collected in future. To formally define R_{L}, we first introduce two intrinsic reward functions of the task to characterize rational, or efficient actions.

$R^{\mathrm{R}}(x, y, a^{\mathrm{R}})$ and $R^{\mathrm{H}}(x, y, a^{\mathrm{H}})$ denote the reward respectively for **R** and **H** taking the action a^{R} or a^{H} in state (x, y). This reward is composed of a high-level reward that promotes similarity towards the goal y and a low-level physical reward associated with the specific action. Hence, actions can have different rewards due to their energy consumption, difficulty, or safety, even if they have the same impact on the similarity towards the task goal.

B. Reward of Legibility

Actions with high rewards defined above are greedily efficient; however, we want **R** also to take inefficient actions that can make the actual goal clear when **H** has a wrong goal expectation. To do that, we derive a reward function of legibility R_{L} for **R** from the principle of rational action, i.e. **H** interprets **R**'s action by assuming **R** is acting efficiently towards the task goal

$$P(a^{\mathrm{R}}|x, y) \propto \exp\left(\beta_1 R^{\mathrm{R}}(x, y, a^{\mathrm{R}})\right), \tag{1}$$

where β_1 is a parameter that **H** assumes how strictly **R** follows the principle of rational action.

Note that the policy for **R** assumed above by **H** is only optimal in one step; for **R** to achieve maximal accumulated reward till the termination, the corresponding POMDP must be solved. However, it is highly unlikely that **H** would have such computational capacity; therefore, we assume that it only considers a greedily efficient policy for **R**.

We assume that **H** does not infer the unknown goal from the whole trajectory at every time step; rather, it infers only based on the current state-action pair and tends to believe what it already believes, which is known as belief perseverance [17] or cognitive inertia [18] in cognitive science. Thus, the system can be represented as a dynamic Bayesian network (DBN) as depicted in Fig 1.

Given the actual task goal y^*, **R** should choose an action a^{R} in state (x, y) that increases the probability $P(y^*|x, y, a^{\mathrm{R}})$ while decreasing $P(y'|x, y, a^{\mathrm{R}}), \forall y' \neq y^*$, yielding a reward function of the form

$$R_{\mathrm{L}}(x, y, a^{\mathrm{R}}) = P(y^*|x, y, a^{\mathrm{R}}) - \lambda \sum_{y' \in \mathcal{Y} \setminus \{y^*\}} P(y'|x, y, a^{\mathrm{R}}), \tag{2}$$

where λ is a tuning parameter that determines how much a wrong expectation should be penalized.

Considering the effect of cognitive inertia that **H** tends to

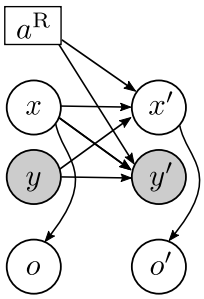


Fig. 1. The system structure as a dynamic Bayesian network. Shaded nodes are partially observable.

believe y' more if $y' = y$, we define

$$P(y'|x, y, a^{\mathrm{R}}) \propto \begin{cases} p_{\mathrm{c}} P(y'|x, a^{\mathrm{R}}), & \text{if } y = y' \\ \frac{1-p_{\mathrm{c}}}{|\mathcal{Y}|-1} P(y'|x, a^{\mathrm{R}}), & \text{otherwise} \end{cases} \tag{3}$$

where $\frac{1}{|\mathcal{Y}|} \leqslant p_{\mathrm{c}} \leqslant 1$ denotes a coefficient indicating how much the human sticks to its previous belief. The probabilities above are computed using Bayes' theorem

$$P(y'|x, a^{\mathrm{R}}) \propto P(a^{\mathrm{R}}|x, y')P(y'|x)\,. \tag{4}$$

C. Goal Inference

The goal inference in HGMDP is achieved by updating the distribution of y at each transition according to

$$b'(y') \propto Z(x', y', a^{\mathrm{R}}, o) \sum_{y} T_{XY}(x, y, a^{\mathrm{R}}, x', y')b(y)\,, \tag{5}$$

where

$$T_{XY}(x, y, a^{\mathrm{R}}, x', y') = T_X(x, y, a^{\mathrm{R}}, x')T_Y(x, y, a^{\mathrm{R}}, x', y') \tag{6}$$

and $Z(x', y', a^{\mathrm{R}}, o) = P(o|x', y', a^{\mathrm{R}})$ is the probability of observing o in state (x', y') after **R** taking action a^{R} in state (x, y).

Recall that we encode the preceding human action in x'; hence, the observation function $Z(x', y', a^{\mathrm{R}}, o)$ is deterministic

$$Z(x', y', a^{\mathrm{R}}, o) = P(o|x', y', a^{\mathrm{R}}) = \begin{cases} 1, & \text{if } o = a^{\mathrm{H}} \\ 0, & \text{otherwise} \end{cases}\,. \tag{7}$$

It can be easily seen from the DBN that x' and y' are conditionally independent given x, y, a^{R}. Thus, we obtain the transition probability

$$T_Y(x, y, a^{\mathrm{R}}, x', y') = P(y'|x, y, a^{\mathrm{R}})\,. \tag{8}$$

Furthermore, we assume that **H** always acts greedily efficiently according to its goal expectation

$$\pi^{\mathrm{H}}(x, y, a^{\mathrm{H}}) \propto \exp\left(\beta_2 R^{\mathrm{H}}(x, y, a^{\mathrm{H}})\right), \tag{9}$$

where β_2 is a parameter that controls how strictly **H** follows the principle of rational action.

Since the uncertainty of the task state transition comes only from **H**, the transition probability is simply

$$T_X(x, y, a^{\mathrm{R}}, x') = P(x'|x, y, a^{\mathrm{R}}) = \pi^{\mathrm{H}}(\tilde{x}, y, a^{\mathrm{H}})\,, \tag{10}$$

where $\tilde{x}$ is the intermediate state reached by **R** taking action a^{R} in state (x, y) and x' by **H** taking action a^{H} in state $(\tilde{x}, y)$.

D. Myopic Heuristic

An optimal legible policy $\pi_{\mathrm{L}}(x, b(y), a^{\mathrm{R}}) \mapsto [0, 1]$ selects actions for **R** to achieve the maximal accumulated reward of legibility. Unfortunately, solving HGMDPs is PSPACE-complete even for deterministic dynamics. Therefore we will not seek exact solutions of this HGMDP; rather, we employ a myopic heuristic to approximate the legible policies. To that end, we first learn the optimal legible policy under each wrong goal expectation of the human collaborator and then switch between those policies according to the current belief state $b(y)$ of the original HGMDP.

When **H** has a fixed wrong goal expectation $y_i \neq y^*$, the HGMDP is reduced to an MDP $M_i = (\mathcal{X}, \mathcal{Y}_i, \mathcal{A}^{\mathrm{R}}, \mathcal{A}^{\mathrm{H}}, T_i, R_{\mathrm{L}}, \gamma)$ where $\mathcal{Y}_i = \{y^*, y_i\}$ and

$$T_i = P(x', y'|x, y, a^{\mathrm{R}}) = \begin{cases} \pi^{\mathrm{H}}(\tilde{x}, y_i, a^{\mathrm{H}}), & \text{if } y' = y_i \\ 0, & \text{if } y' \neq y_i \end{cases}\,, \tag{11}$$

where $\tilde{x}$ is the intermediate task state reached by executing a^{H} in state x.

In defining the reward of legibility, we still assume that **H** will virtually change its mind despite our assumption of fixed wrong goal expectation

$$R_{\mathrm{L},i}(x, y_i, a^{\mathrm{R}}) = P(y^*|x, y_i, a^{\mathrm{R}}) - \lambda P(y_i|x, y_i, a^{\mathrm{R}})\,. \tag{12}$$

We apply a standard Q-learning [19] algorithm to solve M_i associated with each possible wrong goal expectation. An episode of Q-learning terminates when the probability $P(y^*|x, a^{\mathrm{R}})$ exceeds a threshold p_{th} or the actual goal is achieved. Thus, we obtain a legible policy $\hat{\pi}_{\mathrm{L}}(x, y_i, a^{\mathrm{R}})$ for each wrong goal expectation y_i.

Recall that the distribution of Y can be updated by (5) at each time step, which allows us to adjust the policy accordingly. A simple heuristic can be obtained as

$$\pi_{\mathrm{L}}(x, b(y), a^{\mathrm{R}}) = \hat{\pi}_{\mathrm{L}}(x, \operatorname*{argmax}_{y \in \mathcal{Y} \setminus \{y^*\}} b(y), a^{\mathrm{R}})\,. \tag{13}$$

That is, **R** acts under the assumption that **H**'s expectation of the task goal is the one with the highest probability. For general POMDPs, such heuristics suffer from poor performance if the uncertainty is high in the belief state [20], as the robot will not actively take *information gathering actions* on the hidden states. To alleviate this, some algorithms [21], [22] incorporate entropy information in the reward structure to encourage the POMDP agent to take actions that decrease the entropy of the belief state. However, our problem involves a special case that the legible actions are in fact "information gathering" in the sense that they increase the probability of the actual goal being inferred by **H**.

As legible actions can be inefficient, we let the robot switch to the greedily efficient policy once the probability assigned to the actual goal reaches a certain threshold, so as to prevent unnecessarily inefficient actions.

III. FEATURE-BASED STATE ABSTRACTION

In order to alleviate the effect of curse of dimensionality [23], we provide a feature-based state abstraction method for assembly tasks. An assembly task can be seen as a combination of objects at the corresponding positions. We call a correct object-position pair a component c and represent an assembly task as a set of its components $T = \{c_1, c_2, ...\}$. In a nonverbal legible assembly problem, the human collaborator is faced with multiple possible tasks $\mathcal{T} = \{T_1, T_2, ...\}$, from which we obtain the set of all task components $\mathcal{C} = \bigcup_{i=1}^{|\mathcal{T}|} T_i$. For each component c_i, we can find the set of tasks to which it belongs $\mathcal{P}_i = \{T_j | c_i \in T_j\}$, to which we refer as parents of c_i. It is not hard to see that different components can have the same parents, i.e. $\mathcal{P}_m = \mathcal{P}_n$. We define an equivalence relation for such components

$$\mathcal{R} = \Big\{(c_m, c_n) | \mathcal{P}_m = \mathcal{P}_n c_m, c_n \in \mathcal{C}\Big\}. \tag{14}$$

A partition Π of $\mathcal{C}$ can then be obtained as $\Pi = \{[c]_\mathcal{R} | c \in \mathcal{C}\}$ with $[c]_\mathcal{R}$ denoting the equivalence class of c with respect to $\mathcal{R}$ and we call these equivalent classes *subtasks*. For simplicity of notation, we rewrite the equation as

$$\Pi = \{E_i | i \in \{1, 2, 3, ...|\Pi|\}\}, \tag{15}$$

where E_i denotes the subtasks for $i \in \{1, 2, 3, ...|\Pi|\}$.

Given an arbitrary task state $x \in \mathcal{X}$ and its corresponding ongoing task T_x as a set of the components built in state x, we count the number of built components for each subtask and represent the task state with these numbers. Formally, we define the following features

$$\phi_i : x \mapsto |E_i \cap T_x|, \tag{16}$$

where $x \in \mathcal{X}$ and $i \in \{1, 2, 3, ...|\Pi|\}$.

Recall that we define a component as a correct object-position pair. Hence, a missing component can result either from a wrong object or a wrong position besides solely vacancy. We call such wrong object-position pairs *errors* and denote the number of errors by an extra feature $\phi_e(x)$. Here we assume that the number of errors is bounded by a maximal value M_e. Together, the task state can be aggregated to $\mathbb{R}^{|\Pi|+1}$ by a feature function

$$\Phi(x) : x \mapsto \big[\phi_e(x), \phi_1(x), \phi_2(x), \ldots, \phi_{|\Pi|}(x)\big]^\mathsf{T} \tag{17}$$

From the abstract task state, a corresponding abstraction for actions follows naturally: we use a_i and $\bar{a}_i$ to represent actions of increasing and decreasing $\phi_i(x)$ respectively and a_e and $\bar{a}_e$ for making and correcting an error.

IV. EXPERIMENTS

In this section we evaluate the proposed HGMDP in a real Human-Robot Collaboration (HRC)-scenario based on an exemplary dyadic pick-and-place experiment with 10 individual subjects ($\mu_{\text{age}} = 26.47$ years; $\mu_{\text{background}} = 2.3$ on a three-point Likert scale ranging from no to professional robotics background).

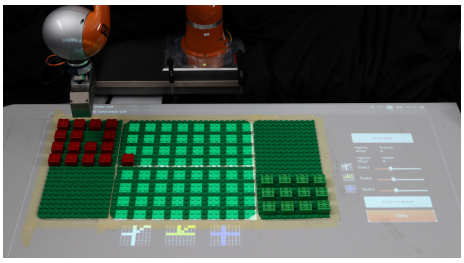
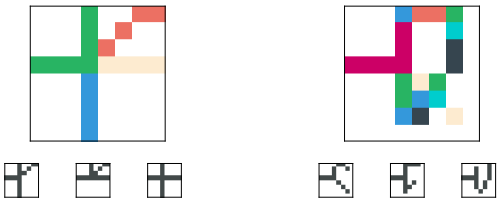


Fig. 2. HRC LEGO-assembly scenario with the goal being unknown to the human collaborator. Participants are asked to give their belief over the possible task goals via the sliding bar on the projected GUI.



(a) Scenario 1 (b) Scenario 2

Fig. 3. Visualization of three pick-and-place goals for the two task scenarios. The subtasks E_i for state abstraction are visualized by color.

A. Experimental Setup

We designed two pick-and-place scenarios in which three different tasks with overlapping subtasks according to (15) are given, as depicted in Fig. 3.

The experimental setup depicted in Fig. 3(b) was characterized by having distinct, overlapping and shared subtasks, whereas in the task scenario shown in Fig. 3(a) no task had a distinct subtask. Each run was assembled by dyads in a round-based manner with the robot acting first. In order to collect consequent user feedback, a GUI was projected upon the workspace from top as shown in Fig. 2, which was used to obtain the human action a^H and self-evaluated belief y over the task goals.

As solely asking for accomplishing the goal would result in barely any difference between the different policies mentioned above, the dyads were asked to assemble the given shape most efficiently, i.e. with the minimum overall travel-distance. This allows the investigation on our claim that a robot can deviate from the efficient policy to decrease the uncertainty of the human collaborator's belief over the task goals.

We compared three robot decision-making modes:

- *efficient* (*E*) In this mode the robot was acting purely efficiently, regardless of the human collaborator's belief, thus assembling the closest component at every step.
- *legible* (*L*) In this mode the HGMDP was applied as outlined in section II.
- *legible with user feedback* (*LF*) In this mode the HGMDP was partially applied. In contrast to *L*, the user-feedback replaced the HGMDP belief estimation.

B. Experimental Procedure

Upon arrival, all participants signed an informed consent form and were surveyed about their background. After this, the experimental setup was explained to the subjects in the form of written text, experimental trials as well as training examples until the subject agreed upon continuation.

Each participant conducted 18 experimental runs such that each decision-making mode was performed 6 times and each scenario 9 times in no particular order. At the end of every assembly task, the participants were asked to answer the questionnaire shown in Table I in a five-point Likert scale. Additionally, the subjects were asked to rate their belief of the task goals after each robot's action via the GUI (Fig. 2).

C. Hypotheses

We propose the following 4 hypotheses upon designing our algorithm to point out the performance and potential:

H1 - *Participants will agree more strongly that the robot's actions are helpful and efficient in mode* L *or* LF *compared to* E. We claim that the efficiency and helpfulness of the robot's actions perceived by the human collaborator is improved by the robot acting efficiently when possible and only selecting legible but inefficient actions when the human collaborator's false belief requires it.

H2 - *Participants will agree more strongly that the robot's actions are responsive in mode* L *or* LF *compared to* E. We claim that the proposed framework allows the robot to adjust its policy according to the inferred goal expectation of the human collaborator, leading to more responsive actions.

H3 - *Participants' belief over the goal will converge faster to the correct goal in mode* L *or* LF *compared to* E. We claim that the legible policies applied by our framework enable the participants to infer the actual task goal more quickly.

H4 - *The overall error rate will be lower in mode* L *or* LF *compared to* E. We claim that an early intervention due to the legible policies helps the human collaborator recover from a wrong belief, thus resulting in lower error rates.

D. Measures and Analysis

The results of the participant surveys are reported in Fig. 5. A Friedmans test for overall comparison was conducted for each question, where the robot decision-making mode is the treatment factor in which we are interested and the task scenario is the blocking factor whose effects need to be taken into account but are not of interest. Post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.017$. The p-values are summarized in Table II.

TABLE I

QUESTIONNAIRE

Q1	*The robot was acting efficiently.*
Q2	*The robot adapted the strategy when I was in doubt about the task.*
Q3	*The robot reacted when I made errors.*
Q4	*The choice of actions of the robot was helpful.*

TABLE II

SUBJECTIVE EVALUATION. EACH CELL HOLDS p-VALUES FOR OVERALL & PAIRWISE COMPARISON. BOLD VALUES ARE STATISTIC SIGNIFICANT.

Question	Overall Comparison	*L* vs *E*	*L* vs *LF*	*E* vs *LF*
Q1	$\mathbf{0.0009}$	$\mathbf{0.0013}$	0.8591	$\mathbf{0.0004}$
Q2	$\mathbf{< 0.0001}$	$\mathbf{< 0.0001}$	0.2789	$\mathbf{0.0002}$
Q3	$\mathbf{< 0.0001}$	$\mathbf{< 0.0001}$	0.5525	$\mathbf{< 0.0001}$
Q4	$\mathbf{< 0.0001}$	$\mathbf{< 0.0001}$	0.8552	$\mathbf{< 0.0001}$

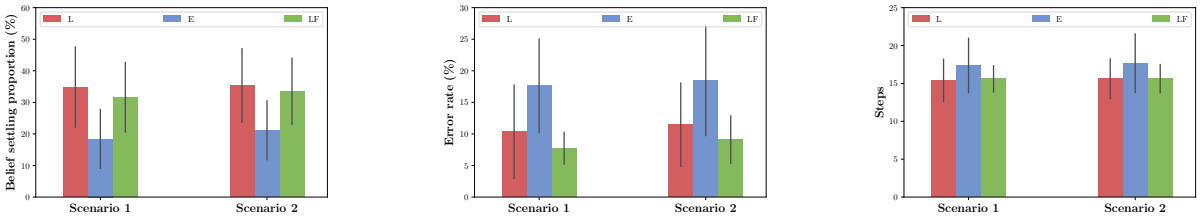
With a statistically significant difference, the participants agreed more strongly that the robot's actions were efficient and helpful in mode *L* or *LF*, compared to *E* (Q1 and Q4). This supports **H1**. Interestingly, we observed a higher variance of the answers for Q1 between the subjects in mode *E*. We attribute this to the possible different definitions of "efficiency" of the participants. While the robot's actions in mode *E* were efficient in terms of the travel-distance, they failed to convey the robot's intention clearly and thus resulted in more steps on average to complete the task, which might be perceived as inefficient by some participants.

Furthermore, the participants agreed more strongly that the robot responded when they were in doubt of the task or made errors in mode *L* or *LF*, compared to *E* (Q2 and Q3). This supports **H2** and suggests that the proposed framework was able to estimate the human collaborator's belief and adjust its policy accordingly. The performance perceived by the participants seems comparable between the mode *L* and *LF*, however. To support this claim, an equivalence test is required in future work.

In order to further evaluate the performance of the proposed framework and the hypotheses mentioned above, we also consider following quantitative measures.

- *Task completion steps* The total number of steps required by the human-robot team to complete the task is measured for all decision-making modes.
- *Belief settling proportion* During the experiment, the participants were asked to give their belief over the task goals after every robot action. We count the steps from the task completion where the human continuously has a correct goal expectation, i.e. the probability assigned to the actual goal is higher than 0.5, and divide it by the total steps of the task and refer it to as the belief settling proportion.
- *Error rate* As a direct measure of a false belief of the human collaborator, the number of errors during the tasks is divided by the number of the task completion steps. We remove the cases across all decision-making modes where the participants guessed the actual goal correctly and thus made no errors.

The quantitative measures show that compared to working with the robot in mode *E*, when the participants were working with the robot in mode *L* or *LF*, they had a larger belief settling proportion (Fig. 4(a)) and lower error rates (Fig. 4(b)) on average, supporting our hypotheses **H3** and **H4**. As shown in Fig. 4(c), the participants also completed the task within fewer steps during the task on average in mode *L* or *LF* compared to *E*. Moreover, we observed that the variance of the task completion steps between the subjects was lower in mode *L* or *LF*, compared to *E*. This can result through the fact that while participants made more errors in mode *E* when they had a wrong goal expectation, there was a certain chance that they guessed the goal correctly from the beginning and thus completed the task within very few steps. As this can happen in the other two modes as well, a lower variance of the task completion steps further



(a) Belief settling proportion (b) Error rate (c) Number of task completion steps

Fig. 4. Mean and standard deviation of the quantitative measures for three robot-decision making modes.

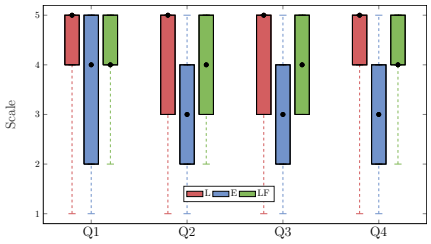


Fig. 5. Answers for each question are grouped by three different modes: *L*, *E* and *LF*. The upper and lower boundaries of the box represent the Interquartile Range (IQR). Whiskers above and below the box indicate the maximum and minimum value of the data. Median is marked by the white circle with a black dot inside.

suggests that the decisions made by the robot in mode *L* and *LF* were more helpful in reducing the potential errors when the human collaborator had a wrong expectation of the task goal, as shown in Fig. 4(b).

V. CONCLUSIONS

In this paper we extend the concept of legibility in motion planning to the domain of sequential decision-making where continuous trajectories are replaced by discrete action-sequences. With one of the major challenges being the human actions as part of the system trajectory, we propose a framework based on hidden goal Markov decision processes (HGMDPs) in which the human collaborator's expectation of the task goal forms the partially observable variable. As solving the resulting HGMDP is PSPACE-complete, policies in reduced fully observable Markov decision processes (MDPs) are obtained offline, and selected according to the online human belief estimation in the original HGMDP.

We evaluate our algorithm through dyadic pick-and-place experiments. In this scenario, the robot deviates from the spatially efficient policy to make the actual task goal more clear according to the estimated human belief. The experimental results confirm the proposed hypotheses with empirical measurements as well as subjective feedback.

Although our general framework is not limited to a specific task setup, the state abstraction method is only applicable for certain assembly scenarios where the potential tasks can be decomposed into object-position pairs. The belief estimation in the HGMDP could be further improved by incorporating richer observations such as eye gaze and hand gestures. Moreover, as the current algorithm only takes into account the selection of abstract actions, future work will consider the integration of legible motion planning into the execution of those abstract actions.

ACKNOWLEDGMENT

This research has been supported by the SIEMENS AG.

REFERENCES

[1] G. Csibra and G. Gergely, "Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans," *Acta Psychologica*, vol. 124, pp. 60–78, 2007.
[2] G. Gergely, Z. Nádasdy, G. Csibra, and S. Bíró, "Taking the intentional stance at 12 months of age," *Cognition*, vol. 56, pp. 165–193, 1995.
[3] K. Kamewari, M. Kato, T. Kanda, H. Ishiguro, and K. Hiraki, "Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion," *Cogn Devel*, vol. 20, pp. 303–320, 2005.
[4] G. Gergely and C. Gergely, "Teleological reasoning in infancy: the naive theory of rational action," *Trends Cogn Sci*, vol. 7, pp. 287–292, 2003.
[5] P. Abbeel and A. Ng, "Apprenticeship Learning via Inverse Reinforcement Learning," in *ICML*, 2004.
[6] B. Ziebart, A. Maas, J. Bagnell, and A. Dey, "Maximum Entropy Inverse Reinforcement Learning," in *AAAI*, 2008, pp. 1433–1438.
[7] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *NIPS*, 2016, pp. 3909–3917.
[8] C. Baker, R. Saxe, and J. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, pp. 329–349, 2009.
[9] A. Dragan, K. Lee, and S. Srinivasa, "Legibility and Predictability of Robot Motion," in *HRI*, 2013.
[10] A. Dragan and S. Srinivasa, "Generating Legible Motion," *Robotics: Science and Systems*, 2013.
[11] F. Stulp, J. Grizou, B. Busch, and M. Lopes, "Facilitating Intention Prediction for Humans by Optimizing Robot Motions," in *IROS*, 2015.
[12] F. Stulp and O. Sigaud, "Policy improvement: Between black-box optimization and episodic reinforcement learning," in *JFPDA*, 2013.
[13] G. Hoffman and C. Breazeal, "Cost-Based Anticipatory Action Selection for Human-Robot Fluency," *IEEE Trans. Robot.*, vol. 23, pp. 952–961, 2007.
[14] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient Model Learning from Joint-Action Demonstrations for Human-Robot Collaborative Tasks," in *HRI*, 2015, pp. 189–196.
[15] A. Fern, S. Natarajan, K. Judah, and P. Tadepalli, "A decision-theoretic model of assistance," *J. Artif. Intell. Res.*, vol. 50, pp. 71–104, 2014.
[16] S. W Ong, S. Wei Png, and D. Hsu Wee Sun Lee, "Planning under Uncertainty for Robotic Tasks with Mixed Observability," *I. J. Robotic Res.*, vol. 29, pp. 1053–1068, 2010.
[17] C. Anderson, M. Lepper, and L. Ross, "Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information," *J Pers Soc Psychol*, vol. 39, pp. 1037–1049, 1980.
[18] G. Hodgkinson, "Cognitive Inertia in a Turbulent Market: the Case of UK Residential Estate Agents," *J. Manag. Stud.*, vol. 34, pp. 921–945, 1997.
[19] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.
[20] D. Aberdeen, "A (revised) survey of approximate methods for solving partially observable markov decision processes," *Nat. ICT Aus., TR*, 2003.
[21] F. Melo and I. Ribeiro, "Transition Entropy in Partially Observable Markov Decision Processes," in *IAS*, 2006, pp. 282–289.
[22] D. Sadigh, S. Sastry, S. Seshia, and A. Dragan, "Information Gathering Actions over Human Internal State," in *IROS*, 2016, pp. 66–73.
[23] R. Bellman, *Dynamic Programming*, 1957.