2017 26th IEEE International Symposium on
Robot and Human Interactive Communication (RO-MAN)
Lisbon, Portugal, Aug 28 - Sept 1, 2017.

Decision-Theoretic Planning Under Uncertainty for Multimodal Human-Robot Interaction

João A. Garcia, Pedro U. Lima and Tiago Veiga

Abstract— **This paper proposes a Decision-Theoretic approach to problems involving interaction between robot systems and human users, which takes into account the latent aspects of Human-Robot interaction, e.g., the user's status. The presented approach is based on the Partially Observable Markov Decision Process framework, which handles uncertainty in planning problems, extended with information rewards to optimize the information-gathering capabilities of the system. The approach is formalized into a framework which considers: observable and latent state variables; gesture and speech observations; and action factors which are related to the agent's actuators or to the information gain goals (Information-Reward actions). Under the proposed framework, the robot system is able to: actively gain information and react according to latent states, inherent to Human-Robot interaction settings; effectively achieve the goals of the task in which the robot is employed; and follow a socially appealing behavior. Finally, the framework was thoroughly tested in a socially assistive scenario, in a realistic apartment testbed and resorting to an autonomous mobile social robot. The experiments' results validate the proposed approach for problems involving robot systems in Human-Robot interaction scenarios.**

I. INTRODUCTION

Social robots need to be capable of developing affective interactions and to empathize with human users [1]. This requirement involves the ability to infer and react according to latent variables: the user's affective and motivational status.

The agent acting in a Human-Robot Interaction (HRI) scenario must take into account the effects of its actions in the human user, which are uncertain, and the sensory information it receives, which is noisy. Planning under these conditions is attainable through Partially Observable Markov Decision Processes (POMDPs) [2]. POMDPs, through the transition and observation models, deal with the aforementioned uncertainty, by probabilistically modeling the possible outcomes of the agent's different actions and the accuracy of the sensory information. Furthermore, the problem of empathizing with the human user adds the goal of information gain on latent (i.e., not directly observed) state variables which is addressed by the extensions to POMDPs introduced by Partially Observable Markov Decision Processes with Information Rewards (POMDPs-IR) [3]. In this context, the paper introduces a POMDP-IR framework for planning under uncertainty in HRI problems, which allows the agent to accomplish a given task, actively infer latent state variables of interest and adapt its behavior accordingly.

The aforementioned framework is implemented in a real robot system, to ensure it is capable of succesfully solving HRI planning problems in practice.

II. RELATED WORK

In HRI scenarios, Decision-Theoretic (DT) approaches to planning based on the POMDP framework are found in assistive scenarios, such as the robot wheelchair [4], where the goal is to recognize the intention of the user but do not include social capabilities to improve recognition. Also, in socially assistive settings, the POMDP framework models the social interaction between robot and human users in, e.g., nursing homes [5], although not taking into account the user's status. Finally, the POMDP was used to model problems with latent variables and adapt the agent's behavior accordingly in an automated hand-washing assistant [6]. However, the agent in the latter work does not actively seek to gain information on the user's status, and is, therefore, limited to react based on a possibly high-uncertainty belief on the hidden variables.

The traditional POMDP model does not allow rewarding low-uncertainty beliefs. Consequently, in order to obtain a certain level of knowledge on the features of interest, the POMDP framework needs to be extended to reward information gain. This extension is provided through the POMDP-IR framework. DT planning based on POMDPs-IR has been applied to the problem of active cooperative perception [3]. The present work, however, is focused on multimodal human-robot interaction.

III. FRAMEWORK DESCRIPTION

A POMDP-IR can be expressed as a tuple $(S, A, T, R, \Omega, O, \gamma)$ where: $S = S_1 \times \cdots \times S_n$ represents the environment's factored state space, defining the model of the world; A is a finite set of actions available to the agent which contains the domain-level action factor A_d and a Information-Reward (IR) action factor A_i for each state factor of interest ($A = A_d \times A_1 \times \cdots \times A_l$, where l is the number of IR actions); T is the transition function that represents the probability of reaching a particular state $s \in S$ by a given state-action pair ($T : S \times A \times S \to [0, 1]$); R is the reward function, which defines the numeric reward given to the agent for each state-action pair ($R : S \times A \to \mathbb{R}$), and is therefore given by $R = R_d(s, a_d) + \sum_{i=1}^{l} R_i(s_i, a_i)$, with $s \in S$, $a_d \in A_d$, $s_i \in S_i$, $a_i \in A_i$, R_d the POMDP reward model and R_i the information reward; Ω is a

All authors are with Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal. This work was partially funded by project CMUP-ERI/HCI/0051/2013 and also by the LARSyS strategic funding from FCT, project UID/EEA/50009/2013.

joao.p.garcia@tecnico.ulisboa.pt
{pal, tsveiga}@isr.tecnico.ulisboa.pt

978-1-5386-3518-6/17/$31.00 ©2017 IEEE

finite set of observations that correspond to features of the environment directly perceived by the agent's sensors; O is the observation function which represents the probability of perceiving observation $o \in \Omega$ after performing action $a \in A$ and reaching state $s' \in S$ ($O : S \times A \times \Omega \rightarrow [0, 1]$); γ is the discount factor, used to weight rewards over time.

The POMDP-IR fits into the classic POMDP framework and can, therefore, be represented as a belief-state Markov Decision Process (MDP), in which the history of executed actions and perceived observations are encoded in a probability distribution over all states: the belief state. Every time the agent performs an action $a \in A$ and observes $o \in \Omega$, the belief is updated by the Bayes' rule:

$$b^{ao}(s') = \frac{P(o|s', a)}{P(o|b, a)} \sum_{s \in S} P(s'|s, a) b(s), \tag{1}$$

where $P(s'|s, a)$ and $P(o|s', a)$ are defined by the Transition and Observation model, respectively, and

$$P(o|b, a) = \sum_{s' \in S} P(o|s', a) \sum_{s \in S} P(s'|s, a) b(s) \tag{2}$$

is a normalizing constant. Furthermore, the value function $V^{\pi}(b)$, defined as the expected future discounted reward given to the agent by following policy π, starting from belief b:

$$V^{\pi}(b) = \mathbf{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(b_t, \pi(b_t)) \Big| b_0 = b\right], \tag{3}$$

where $R(b_t, \pi(b_t)) = \sum_{s \in S} R\big(s, \pi(b_t)\big) b_t(s)$, remains approximately Piecewise Linear Convex (PWLC) in the POMDP-IR framework. This way, the most common algorithms for solving POMDPs, which exploit the PWLC representation of the value function, can also be used to solve POMDPs-IR. The optimal policy π^* is characterized by the optimal value function V^*, that statisfies the Bellman optimality equation:

$$V^*(b) = \max_{a \in A}\left[R(b, a) + \gamma \sum_{o \in O} P(o|b, a) V^*(b^{ao})\right]. \tag{4}$$

Figure 1 represents the projected POMDP-IR framework for multimodal HRI, as a two-stage Dynamic Bayesian Network (DBN), which depicts the dynamics of the HRI problem.

A. States and Transitions

The agent acting in a HRI scenario considers two types of state factors: the *task* variables T and the *person* variables P. The *task* variables model the environment features that provide information on the progress of the tasks. On the other hand, the *person* variables track the human state and are inherently latent. The latter are used to gain information on the human user's affective and motivational status and adapt the robot behavior accordingly.

The number of state variables depend on the amount of features essential to represent the environment and is, therefore, dependent on the specific task. The criteria for the selection of states involve a trade-off between operational complexity and predicted system performance, since operational complexity increases with the number of states.

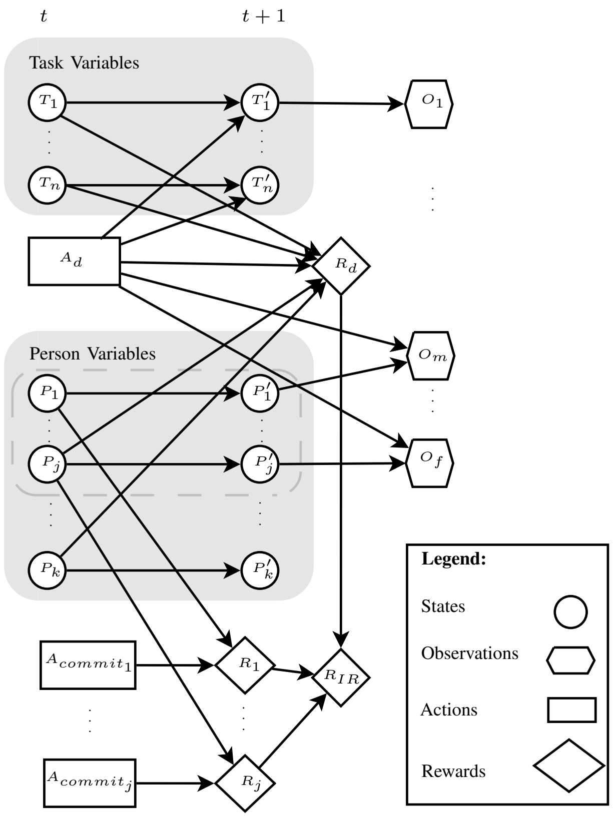


Fig. 1. DBN representation of the DT model for multimodal HRI.

Furthermore, depending on the objectives of the agent acting in a HRI setting, the *task* variables might not exist. This is the case when the single goal of the agent is to gain information on the human user.

A *person* variable can have a constant value over time if its value does not change during the task. This is the case of personal traits (e.g., *Personality* and *Preferences*), which are relevant for the robot behavior and do not change for the duration of the interaction. In Figure 1, P_k represents a constant *person* variable.

Otherwise, *person* variables are inferred from the user's behavior at each time step (factors P_1 to P_j in Figure 1), which is represented in the model's observations. These state variables may consist of state factors of interest, according to the POMDP-IR framework.

B. Observations and Observation Model

In a social HRI setting, observations reflect the user's behavior. This behavior is used to monitor the progress of the task and infer the user's affective and motivational status.

Observations are discrete, symbolic values, classified from sensory data, which correspond to features of the environment that are observable in a given state.

The observation factors are contingent on the sensory capabilities of the robot system. Nevertheless, the correct understanding of the user's status relies on the agent being capable of recognizing human communication methods.

Consequently, the robot system ought to be able to recognize speech and gestures in order to understand the human user's affective and motivational status.

The observation model is of key importance in the achievement of the information gain goals of the agent. It reflects the probability of receiving a certain observation, given the state of the environment and the action performed. Certain actions, such as questioning or approaching the user, increase the probability of perceiving certain observations. This fact is of utter importance to actively gain information on the user's status. The dependency on the action is represented in observations O_m to O_f in Figure 1.

C. Actions

The model of Figure 1 comprehends two types of actions: A_d and A_{commit}. The former have an effect on the environment and are dependent on the actuators of the agent, while the latter are used for the information gain goals of the agent.

Typically, the action domain A_d contains the minimum set of functionalities which allow the agent to complete its tasks. Social robots need to communicate in a natural, easily understandable way with the human users. To achieve this objective, the robot must be able to express different moods and emotions. Consequently, the action domain A_d of a social robot ought to include speech and/or gestural capabilities and/or graphical emotion displays.

Following the POMDP-IR framework, besides the domain-level action factor A_d, the model has additional action factors A_{commit} for each state factor of interest. The state factors of interest, in the problem under study, are included in the *person* variables, as these contain the aforementioned affective and motivational state of the human user. The actions A_{commit} allow rewarding the agent for decreasing the uncertainty regarding particular features of the environment.

D. Reward Model

In the DT model of Figure 1, rewards are either associated with task objectives: R_d, or with the information gain goals: $R_i, i = 1, \ldots, j$. The sum of these rewards, R_{IR}, constitute the reward awarded to the agent at each time step.

The behavior of the robot consists of the sequence of domain actions A_d the agent performs. In the social HRI scenario, and in order to adapt the robot's behavior to the user's affective and motivational status, the reward assigned to an action depends not only on the *task* variables but also on the *person* variables.

The information rewards R_i influence the behavior of the agent, with the purpose of achieving a low uncertainty regarding certain *person* variables. The value of these rewards are dependent on the threshold of knowledge required, according to the POMDP-IR framework [3].

IV. SELECTED APPLICATION

The proposed approach was tested in a case study which considers an active rehabilitation therapy task. In this task, the patient moves the affected limb by him/herself, while the robot therapist has the functions of coaching and motivating.

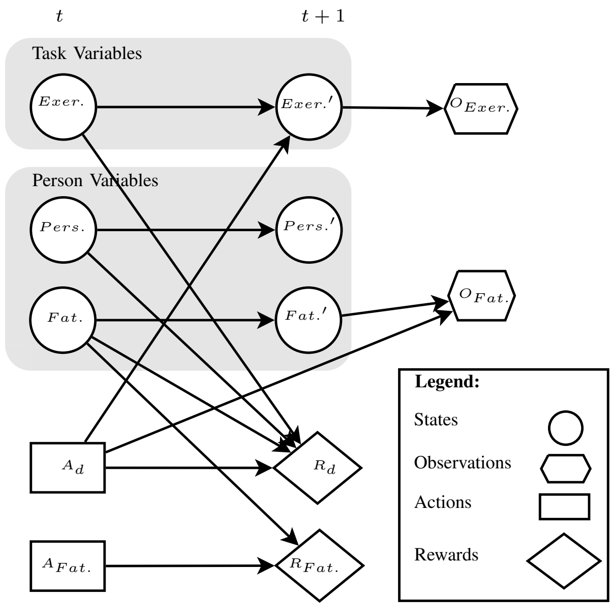


Fig. 2. DBN representation of the DT model for the robot therapist.

Summarizing, the goals of the robot therapist in the considered rehabilitation scenario are: to help the user in the given setting, by monitoring the patient's movements (e.g., encourages the patient to continue if he/she stops performing the exercise); and to adapt its behavior and, consequently, the therapy style (e.g., nurture or challenge the patient), in accordance with the patient's affective and motivational status.

A. Decision-Theoretic Model for the Robot Therapist

The application of the proposed framework to the robot therapist scenario results in the DT model represented in Figure 2.

1) States: The significant features of the environment in which the robot is to operate are related to the human user. The fulfillment of the task's objectives require that the agent keeps track of the user's movements (state $Exer.$), possesses knowledge regarding relevant personal traits ($Pers.$) of the user and infers his/hers affective status ($Fat.$). Therefore, the proposed DT model considers the state space represented, in factored form, in Table I.

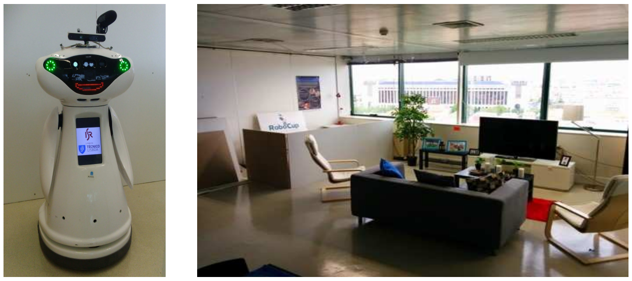
2) Observations: The observation space is represented, in factored form, in Table I. Observations reflect the relevant behavior of the patient, in accordance with the task's goals. In the present case study, the agent ought to classify the movement performed by the patient ($O_{Exer.}$) and his/hers affective status ($O_{Fat.}$). $O_{Exer.}$ is obtained by visual classification of the patient's gestures and $O_{Fat.}$ through classification of the user's verbal responses.

3) Actions: The proposed DT model considers two action factors: the *Action Domain* A_d and the *IR Action* $A_{Fat.}$. At each time step, the agent chooses one value for each action factor. The possible values for the action factors are represented in Table I.

TABLE I

STATE, OBSERVATION AND ACTION SPACES FOR THE ROBOT THERAPIST CASE STUDY

	Factors	Values
States	$Exer.$	Correct, Incorrect
	$Pers.$	Introverted, Extroverted
	$Fat.$	Tired, Energized
Observations	$O_{Exer.}$	Proper, Wrong
	$O_{Fat.}$	Weary, Energetic, None
Actions	A_d	Nurture, Challenge, Query Patient, End Therapy, None
	$A_{Fat.}$	Commit Tired, Commit Energized, Null



(a) Robot Platform used in the experiments.

(b) Living room area of the ISRoboNet@Home testbed.

Fig. 3. Experimental setup for the robot therapist case study.

4) Transition, Observation and Reward Functions: The proposed framework allows to take into account the effects of time in the states of the DT model. Namely, in the current case study, the transition function T encodes that $b(Fat. = Tired)$ increases at each time step in the absence of opposing observations ($O_{Fat.} = Energetic$). That is, the agent realistically believes that the patient is feeling more tired over time. The transition function of this case study dictates that the probability of the patient correctly performing the exercise ($Exer. = Correct$) increases with the motivation actions ($Nurture$ or $Challenge$). Moreover, Personality ($Pers.$) is modeled as a constant variable, not inferred by the agent, as its value does not change during the task.

The observation function O encodes the error in sensory data classification. This means, for instance, that even if the patient's gesture is classified as incorrect ($O_{Exer.} = Wrong$), the agent's belief on $Exer. = Incorrect$ is not 100% and the robot might require more information before motivating the patient. Furthermore, the probabilities in O take into account that information-gathering actions (such as $Query\,Patient$) increase the probability of perceiving a verbal response from the user (e.g., $O_{Fat.} = Weary$).

The DT model of Figure 2 rewards IR actions ($R_{Fat.}$) and A_d actions (R_d). The information rewards are defined, in accordance with the POMDP-IR framework, so that the agent actively seeks to have a certainty on $Fat.$ greater than 75% (i.e., $b(Fat. = Tired) > 0.75$ or $b(Fat. = Energized) > 0.75$). Actions in A_d are rewarded in accordance with the state of the environment: *Encouragement* actions ($Nurture$ and $Challenge$) are rewarded 0.2 whenever the patient is incorrectly performing the exercise or 0.1 when he/she shows signs of feeling tired, and penalized -0.1 otherwise. The reward given to each action also depends on the state factor $Pers.$: for an *Introverted* person, the $Nurture$ action is preferred while the $Challenge$ action is favored for an *Extroverted* person; The $Query\,Patient$ action is penalized with -0.2; $None$ is not rewarded nor penalized; $EndTherapy$ receives high penalization (-1) when the patient feels energetic and a reward of 0.1 otherwise. Rewards are defined over the abstract states and actions of the DT model. Thus, because it would be very impractical to obtain the models from empirical studies, especially as the system becomes more complex, the aforementioned reward values are tuned to lead to a policy which handles the different patients adequately. The discount factor in this case study is $\gamma = 0.9$.

V. EXPERIMENTS

The robot therapist case study was implemented in a networked robot system which consists of: the MOnarCH robot platform, represented in Figure 3(a) and an external Kinect camera; and interacted, in four different experiments, with distinct persons, in in the ISRobotNet@Home Testbed[1], which is represented in Figure 3(b).

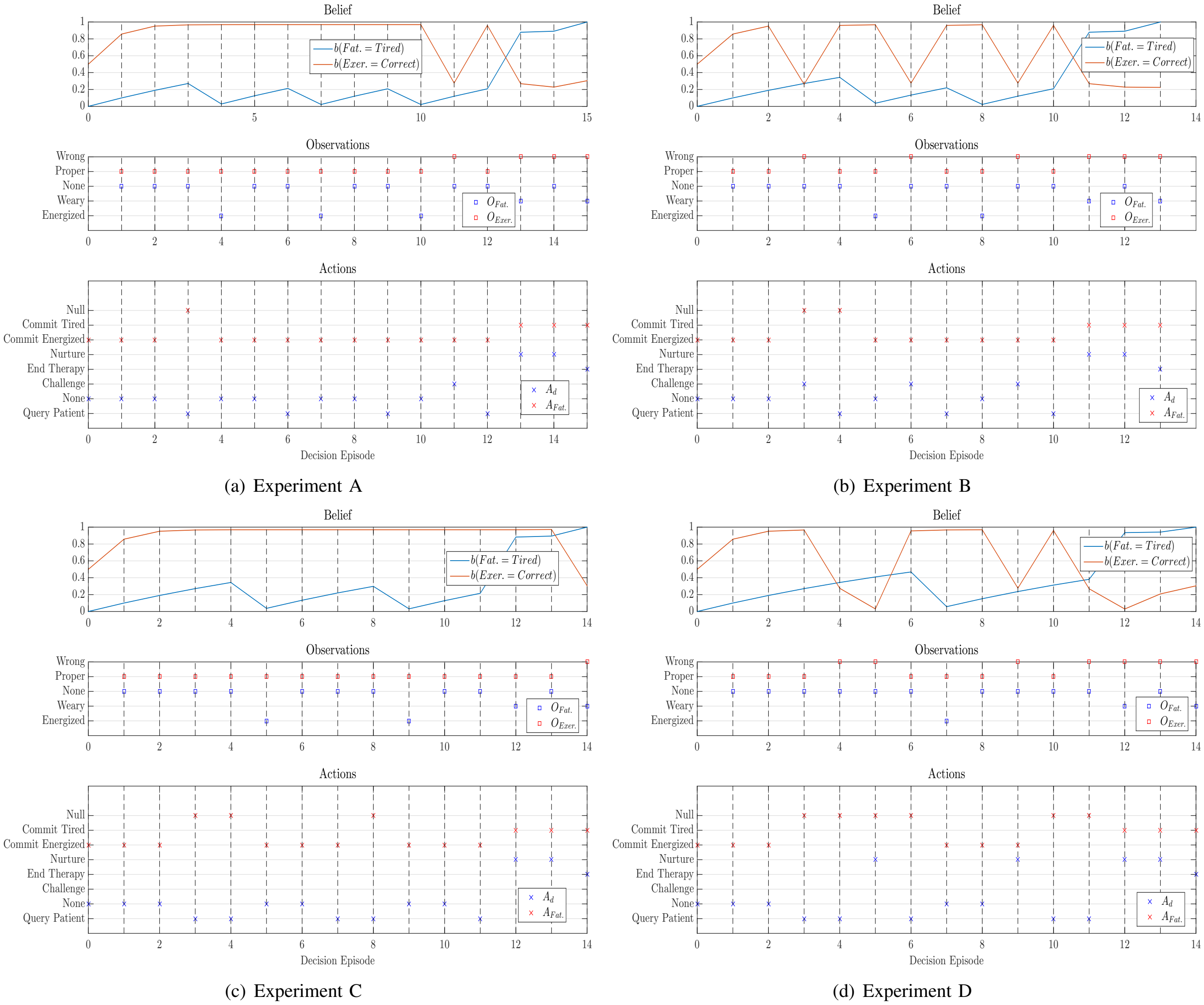
A. Experimental Results

Each experiment considers a different user, which is classified according to his/hers personality (i.e., as introverted or extroverted), and with regard to his/hers ability to perform the exercise (athletic or unfit). The experiments carried out within this work were recorded and are available at `https://goo.gl/TlyXGT`. Figure 4 plots the data acquired in the experiments, namely the observations, actions and belief on the two key state factors considered: $Fat.$ and $Exer.$. Figure 5 represents an episode of experiment B where the robot interacts with the user.

1) Experiment A: This experiment considers a user which is classified as extroverted ($Pers. = Extroverted$) and athletic. The user feels energetic for the first fifty seconds (decision step 10), approximately, and tired afterwards.

At the beginning, the robot chooses not to act, since the exercise is well performed and the agent has a low uncertainty regarding the *fatigue* status of the user. This uncertainty on the state factor $Fat.$, however, increases over time, driving the robot to actively seek to reduce it, by querying the user (decision step 3). The answer ($O_{Fat.} = Energetic$), informs the robot that the user is still active and motivated, increasing the certainty on $Fat. = Energized$. This behavior is repeated until the user does not perform correctly the exercise ($O_{Exer.} = Incorrect$) in decision step 11. Then, the robot motivates the person through a challenging approach due to the considered *personality* of the user and the current *fatigue* status. After receiving information that the user now feels tired ($O_{Fat.} = Weary$), the robot changes therapy style and

[1] `http://welcome.isr.tecnico.ulisboa.pt/isrobonet/`



(a) Experiment A

(b) Experiment B

(c) Experiment C

(d) Experiment D

Fig. 4. Evolution of the Belief on the states $Fat.$ and $Exer.$ w.r.t. the decision episode, the observations received and the actions performed, for each experiment.

adopts a nurturing approach. As the user continuously shows not being able to carry out the exercise and the certainty on $Fat. = Tired$ increases, the robot finally chooses to end the therapy in decision step 15.

2) Experiment B: This experiment considers a user classified as extroverted ($Pers. = Extroverted$) and unfit. The user feels energetic for the first forty seconds, approximately, and tired afterwards.

The behavior of the robot is similar to the previous experiment while the user shows feeling energetic and correctly performs the exercise. Nonetheless, the user incorrectly performs the exercise more often, at which occasions the robot acts in motivating with a challenging approach, while the agent believes the user feels motivated/energetic. Even though motivating the user, the robot keeps track of his/hers *fatigue* and reacts when the uncertainty on $Fat.$ is high. Finally, the agent ends the therapy once it persistently observes the user is not performing the exercise and feels tired.

3) Experiment C: This experiment considers a user classified as introverted ($Pers. = Introverted$) and athletic. The patient feels energetic up to, approximately, 45 seconds (decision step 9), and tired afterwards.

The behavior of the robot is heavily dependent on its knowledge regarding the fatigue status of the user. While the uncertainty on the $Fat.$ state factor is high, the robot queries the user. Since the uncertainty on $Fat.$ increases over time, the agent performs the action $Query\ Patient$ until it perceives an answer $O_{Fat} = Energetic$ or $O_{Fat} = Weary$ (decision steps 3 & 4 / 7 & 8). Nevertheless, the robot performs the therapy task while actively gathering information on the environment, motivating the user once the belief on $b(Fat. = Tired)$ is high, and ending the therapy appropriately.

4) Experiment D: This experiment considers a user which is classified as introverted ($Pers. = Introverted$) and unfit.

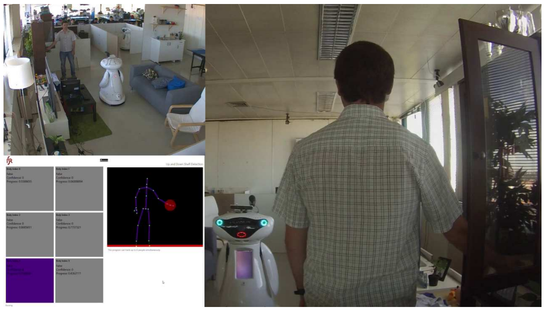


Fig. 5. Episode of the experiment B where the robot queries the user. Right and top left images show different views of the interaction between the robot and the human; Bottom left image represents the interface of the gesture classification application.

The user feels energetic for the first 40 seconds (decision step 8), approximately, and tired onward.

The behavior of the robot changes in accordance with its belief on the states of the environment. In the present experiment, there is a "trade-off" between motivating or querying the user depending on the belief over the state factors $Fat.$ and $Exer.$. In decision step 3, the agent queries the agent due to the high uncertainty on $Fat.$. Afterwards, the agent perceives no answer but observes the user incorrectly performed the movement. This observation does not translate, however, into an absolute certainty on the exercise having been incorrectly performed ($b_4(Exer. = Correct) \approx 0.3$), since the DT framework takes into account sensor related noise. The agent, then, queries the user once again (decision step 4), due to the increasing uncertainty on the *fatigue* of the user. Once again, the Network Robot System (NRS) receives no answer ($O_{Fat.} = None$), and observes the user incorrectly performed the movement. This time, the agent's belief on $Exer. = Incorrect$ is higher ($b_5(Exer. = Incorrect) \approx 0.95$) and it motivated the user. Nevertheless, the uncertainty on $Fat.$ is still high on decision step 6 and the robot once again queries the user, perceiving this time an answer.

For the rest of the experiment, the robot follows a behavior similar to the previous experiments, until it ends the trial in decision step 14.

B. Discussion

Table II details the behavior of the robot for each experiment. As expected: the number of motivation actions is higher for the users classified as unfit, which incorrectly perform the exercise more often than the athletic users; and the number of query actions is higher for the users classified as introverted.

The robot detected the fatigue status change from $Energized$ to $Tired$ in all the experiments. Moreover, the agent motivated the user upon detection of faulty movements, either immediately after observing $O_{Exer.} = Wrong$ (experiments A, B and C) or after two consecutive observations (experiment D). Finally, the agent ended the therapy when consistently observing the user was not capable of proceeding with the exercise.

TABLE II
BEHAVIOR OF THE ROBOT WITH REGARD TO THE EXPERIMENT

	A	**B**	**C**	**D**
Motivation actions	3	5	2	4
Query actions	4	3	5	5
Time elapsed until agent detected change of users status (s)	15	15	15	20
Time elapsed until agent ends therapy since it detected user is tired (s)	10	10	10	10
Duration of the experiment (s)	75	65	70	70

Overall, the DT approach to planning in the robot therapist resulted in a behavior capable of achieving the task and information goals, adaptive to the user's status and socially appealing.

VI. CONCLUSION

Building on the POMDP-IR framework, this work introduced a DT approach to planning under uncertainty with information rewards in social HRI. The properties of the DT framework were demonstrated in the robot therapist case study and the experiments' results validate the proposed framework for a problem involving robot systems in HRI scenarios.

In future work and to further validate the framework developed within this work, further experiments ought to be performed, considering distinct scenarios of HRI. Furthermore, the model ought to be estimated from experimental data [7] or learned through Reinforcement Learning (RL) approaches [8], to overcome practical issues inherent to the implementation of MDP-based models.

REFERENCES

[1] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[2] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99 – 134, 1998.

[3] M. T. J. Spaan, T. S. Veiga, and P. U. Lima, "Decision-theoretic planning under uncertainty with information rewards for active cooperative perception," *Autonomous Agents and Multi-Agent Systems*, vol. 29, no. 6, pp. 1157–1185, 2015.

[4] T. Taha, J. V. Miro, and G. Dissanayake, "Pomdp-based long-term user intention prediction for wheelchair navigation," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, May 2008, pp. 3920–3925.

[5] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Special issue on Socially Interactive Robots, Robotics and Autonomous Systems*, vol. 42, no. 3 - 4, pp. 271 – 281, 2003.

[6] J. Hoey, P. Poupart, A. v. Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis, "Automated Handwashing Assistance for Persons with Dementia Using Video and a Partially Observable Markov Decision Process," *Computer Vision and Image Understanding*, vol. 114, no. 5, pp. 503–519, May 2010.

[7] S. Koenig and R. Simmons, "Xavier: A robot navigation architecture based on partially observable markov decision process models," in *Artificial Intelligence Based Mobile Robotics: Case Studies of Successful Robot Systems*, R. B. D. Kortenkamp and R. Murphy, Eds. MIT Press, 1998, pp. 91 – 122.

[8] T. Jaakkola, S. P. Singh, and M. I. Jordan, "Reinforcement learning algorithm for partially observable markov decision problems," in *Advances in Neural Information Processing Systems 7*. MIT Press, 1995, pp. 345–352.