Robot Fast Adaptation to Changes in Human Engagement During Simulated Dynamic Social Interaction With Active Exploration in Parameterized Reinforcement Learning

Mehdi Khamassi, George Velentzas, Theodore Tsitsimis, Costas Tzafestas

Robot Fast Adaptation to Changes in Human Engagement During Simulated Dynamic Social Interaction With Active Exploration in Parameterized Reinforcement Learning

Mehdi Khamassi, George Velentzas, Theodore Tsitsimis, and Costas Tzafestas, *Member, IEEE*

***Abstract*—Dynamic uncontrolled human–robot interactions (HRIs) require robots to be able to adapt to changes in the human's behavior and intentions. Among relevant signals, nonverbal cues such as the human's gaze can provide the robot with important information about the human's current engagement in the task, and whether the robot should continue its current behavior or not. However, robot reinforcement learning (RL) abilities to adapt to these nonverbal cues are still underdeveloped. Here, we propose an active exploration algorithm for RL during HRI where the reward function is the weighted sum of the human's current engagement and variations of this engagement. We use a parameterized action space where a meta-learning algorithm is applied to simultaneously tune the exploration in discrete action space (e.g., moving an object) and in the space of continuous characteristics of movement (e.g., velocity, direction, strength, and expressivity). We first show that this algorithm reaches state-of-the-art performance in the nonstationary multiarmed bandit paradigm. We then apply it to a simulated HRI task, and show that it outperforms continuous parameterized RL with either passive or active exploration based on different existing methods. We finally test the performance in a more realistic test of the same HRI task, where a practical approach is followed to estimate human engagement through visual cues of the head pose. The algorithm can detect and adapt to perturbations in human engagement with different durations. Altogether, these results suggest a novel efficient and robust framework for robot learning during dynamic HRI scenarios.**

Manuscript received June 30, 2017; revised December 30, 2017 and April 18, 2018; accepted May 17, 2018. Date of publication July 5, 2018; date of current version December 7, 2018. The work of M. Khamassi was supported in part by EU-funded Project BabyRobot (H2020-ICT-24-2015) under Grant 687831, in part by the Agence Nationale de la Recherche (ANR-12-CORD-0030 Roboergosum Project and ANR-11-IDEX-0004-02 Sorbonne-Universités SU-15-R-PERSU-14 Robot Parallearning Project), in part by Labex SMART (ANR-11-LABX-65 Online Budgeted Learning Project), and in part by the Centre National de la Recherche Scientifique (Mission pour l'Interdisciplinarité ROBAUTISTE Project and PICS 279521 SocialRobot Project). The work of C. Tzafestas was supported by EU-funded Project BabyRobot (H2020-ICT-24-2015) under Grant 687831.
(Corresponding author: Mehdi Khamassi.)

M. Khamassi is with the School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece, and also with the Institute of Intelligent Systems and Robotics, Sorbonne Université, CNRS, 75005 Paris, France (e-mail: mehdi.khamassi@upmc.fr).

G. Velentzas, T. Tsitsimis, and C. Tzafestas are with the School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece (e-mail: ktzaf@cs.ntua.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2843122

***Index Terms*—Active exploration, bandits, human–robot interaction (HRI), meta-learning, reinforcement learning (RL).**

I. Introduction

Developing social robots dedicated to interacting and cooperating with humans requires endowing robots with learning capabilities that enable them to adapt quickly and on the fly to changes in humans' behavior and intentions. While one of the explored ways to achieve this has been through the study of robot verbal communication abilities [1]–[6], nonverbal cues such as the human's gaze can provide the robot with important information about the human's current engagement in the task [7], and whether the robot should continue its current behavior or not. Indeed, primates naturally and implicitly monitor mutual gaze and gaze following behaviors to evaluate the level of joint attention during social interaction, and to establish common ground for efficient joint action [8].

Researches in the field of social robotics have recently shown a growing interest in monitoring human and robot gaze during social interaction [9]–[12]. Results show that gaze following improves intention readout, efficiency of joint action, and arouses on human partners the illusion of a social intelligence. Conversely, it has been proposed that monitoring the level of engagement of the human during the task, for instance through the monitoring of body posture and gaze, may provide the robot with crucial information to assess how it is perceived by the human, how this perception changes according to the behaviors shown by the social robot, and hence to improve the quality of human–robot interaction (HRI) [6], [7], [13]–[15]. According to [16], "Engagement is a category of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control." However, to our knowledge no one has yet proposed a way to make the robot learn on the fly in response to changes in human engagement. More generally, robot reinforcement learning (RL) abilities based on nonverbal cues during HRI are still largely underdeveloped, mostly due to the high level of unpredictability and variability of human behavior, but also due to the difficulty in coping with the high-dimensional continuous action space available to the robot during such scenarios. Some studies have previously applied RL techniques for robot adaptation during interaction (e.g., [17]). However, this was made possible through the prior categorization of a small number of discrete stimuli and actions that the robot had to deal with, which prevents generalization to more complex tasks requiring continuous motor actions.

The original contribution of this paper consists of several aspects. To our knowledge, this paper constitutes the first proposal to use human engagement monitoring signals as a reward signal for robot RL during nonverbal social interaction. Here the proposed reward function consists in a weighted sum of the human's current engagement and variations of this engagement (so that a low but increasing engagement is rewarding). Second, this paper proposes a way to apply a *parameterized* version of RL [18], [19] to HRI: this employs a set of discrete actions $A_d = \{a_1, a_2, \ldots, a_k\}$, where each action $a \in A_d$ features m_a continuous parameters $\{\theta_1^a, \ldots, \theta_{m_a}^a\} \in \mathbb{R}^{m_a}$, which enables to benefit from the simplicity of task decomposition into a small set of discrete actions while at the same time being able to exploit the precision of continuous motor execution. We finally propose a way to achieve robot fast adaptation during social interaction through active exploration [20]–[24]. The proposed solution relies on a novel combination of existing methods applied to a simple HRI scenario in the following manner. We apply Gaussian exploration [25] to actions' continuous action parameters, which in the original formulation uses a fixed Gaussian width σ, hence, a fixed exploration rate. Here, we apply a noiseless version of the meta-learning algorithm of [26], which tracks online variations of the agent's performance measured by short-term and long-term reward running averages. At each timestep, we use the difference between the two averages to simultaneously tune the inverse temperature β_t used for selecting between discrete actions a_j, and the width σ_t of the Gaussian distribution from which each continuous action parameter θ_i^a is sampled around its current value.

The rest of this paper is organized as follows. In the next section, we present the detailed formulation of the algorithm. We then present a series of numerical experiments to test it. We first simulate a standard nonstationary (i.e., switching) multi-armed bandit paradigm proposed by [27]. We show that the algorithm reaches similar performance to one of the state-of-the-art upper confidence bound algorithms, while also being generalizable to continuous actions and multistep tasks (which is not the case for bandit methods). We then apply the proposed algorithm to a simple simulated HRI task, where the algorithm tries to maximize reward computed as an approximate and partial measure of engagement of the human in the task, this engagement representing the attention that the human pays to the robot and its actions. We show that the proposed algorithm outperforms continuous parameterized RL both without active exploration and with active exploration based on different existing methods: uncertainty variations measured by a Kalman-RL algorithm [28], exploration bonuses based on computational neuroscience methods [29], [30]. Finally, we test the performance of the algorithm in a more realistic version of the HRI task, where a practical approach is followed to estimate human engagement through visual cues of the head pose. We then measure the adaptation of the algorithm to engagement perturbations simulated as changes in the optimal action parameter and we quantify its performance for variations in perturbation duration and measurement noise, thus illustrating the robustness of the algorithm.

A preliminary version of this work has been presented at a conference [31] and at a workshop [32]. Nevertheless, this paper includes more comparisons with alternative algorithms, includes a novel exhaustive parameter search for each tested algorithm, uses a different method for the engagement estimation process which gives better results than in [32], and presents an extended description and discussion of the work.

II. Active Exploration Algorithm

This section describes the mathematical formulation underlying the proposed active exploration method. The proposed meta-learning algorithm is then summarized at the end of the section (Algorithm 1). It first employs Q-learning [33] to learn the value of discrete action $a_t \in A_d$ selected at timestep t in state s_t

$$\delta_t = r_t + \gamma \max_a (Q_t(s_{t+1}, a)) - Q_t(s_t, a_t) \tag{1}$$

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_Q \delta_t \tag{2}$$

where α_Q is a learning rate and γ is a discount factor. The probability of executing discrete action a_j at timestep t is given by a Boltzmann softmax equation

$$P\left(a_j | s_t, \beta_t\right) = \frac{\exp\left(\beta_t Q_t\left(s_t, a_j\right)\right)}{\sum_a \exp(\beta_t Q_t(s_t, a))} \tag{3}$$

where β_t is a dynamic inverse temperature meta-parameter which will be tuned through meta-learning (see below).

In parallel, continuous parameters $\widetilde{\theta}_{i,t}^{a_j}$ with which action a_j is executed at timestep t are selected from a Gaussian exploration function centered at the current values $\theta_{i,t}^{a_j}(s_t)$ in state s_t of the parameters of this action [25]

$$P\left(\widetilde{\theta}_{i,t}^{a_j} | s_t, a_j, \sigma_t\right) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\left(\widetilde{\theta}_{i,t}^{a_j} - \theta_{i,t}^{a_j}(s_t)\right)^2 / \left(2\sigma_t^2\right)\right) \tag{4}$$

where the width σ_t of the Gaussian is a meta-parameter which will be tuned through meta-learning (see below) and action parameters $\theta_{i,t}^a(s_t)$ are learned with a continuous actor–critic algorithm [25]. A reward prediction error is computed from the critic: $\delta_t = r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$ and is used to update the parameter vectors ω_t^C and ω_t^A of the neural network function approximations in the critic and the actor

$$\omega_{i,t+1}^C = \omega_{i,t}^C + \alpha_C \delta_t \frac{\partial V_t(s_t)}{\partial \omega_{i,t}^C} \tag{5}$$

$$\omega_{i,t+1}^A = \omega_{i,t}^A + \alpha_A \delta_t \left(\widetilde{\theta}_{i,t}^a - \theta_{i,t}^a(s_t)\right) \frac{\partial \theta_{i,t}^a(s_t)}{\partial \omega_{i,t}^A} \tag{6}$$

where α_C and α_A are learning rates. In contrast to the original version where ω_t^A updates are performed only when $\delta_t > 0$ [25]—which occasionally led to divergence in our simulations—here we update them all the time and proportionally to δ_t as in [34].

Finally, in order to perform active exploration, we need to dynamically update β_t and σ_t through a meta-learning process based on variations of the robot's performance. The idea is that increases in the average reward obtained by the robot can be interpreted as improvement of performance which can thus result in increasing the exploitation of learned action values [26], [35]. Conversely, drops in the average reward can be interpreted as signs of a change in the task conditions and thus as a need to re-explore. Nevertheless, the average reward is not an absolute measure and should rather be considered relatively to a reference such as the estimated average value of the task [36]. For instance, in tasks where only punishments are received, the average value of the task is negative, but should not be interpreted as an indication that the robot should only explore and never exploit. Thus here, following the proposition of [26], we measure a long-term reward running average $\bar{\bar{r}}_t$ serving as reference, and a short-term one $\bar{r}_t$ serving as current measure of performance. When $\bar{r}_t > \bar{\bar{r}}_t$, this means that the current performance is above average and that exploration can be decreased. When $\bar{r}_t < \bar{\bar{r}}_t$, this means that the current performance is below average and that exploration should be increased. Contrary to the noisy version of [26] which can lead to meta-learning instability, here we implement a noiseless version of the algorithm. We compute short- and long-term reward running averages in the following manner:

$$\Delta\bar{r}_t = (r_t - \bar{r}_t)/\tau_1 \text{ and } \Delta\bar{\bar{r}}_t = \left(\bar{r}_t - \bar{\bar{r}}_t\right)/\tau_2 \tag{7}$$

where τ_1 and τ_2 are two time constants. We then update β_t and σ_t with

$$\beta_{t+1} = (\mathcal{R} \circ \mathcal{F})\left(\beta_t, \mu\tau_2\Delta\bar{\bar{r}}_t\right) \quad \text{and} \quad \sigma_{t+1} = \mathcal{G}\left(\mu\tau_2\Delta\bar{\bar{r}}_t\right) \tag{8}$$

where $\mathcal{R}(x)$ is a rectifier function, $\mathcal{F}(x, y)$ is an affine function, μ is a learning rate and $0 < \mathcal{G}(x) < 0.1M$ is a sigmoid function, with M denoting the parameter range.

We also compared this meta-learning algorithm with the Kalman Q-learning proposed by [28]. We first tested the original formulation which proposes a purely exploratory agent by replacing Q-values in (3) by the action-specific diagonal terms of the covariance matrix—these terms representing the current variance/uncertainty about an action's Q-value. We then tested an extended version of the algorithm were diagonal terms of the covariance matrix are treated as *exploration bonuses* b_t^a which, like in a previous computational neuroscience work [29], are multiplied by a weight η and added to Q-values in (3). A particular novelty here is that we also use the covariance terms b_t^a in replacement of $\bar{\bar{r}}_t$ in (8) to tune action-specific σ_t^a with function $\mathcal{G}(x)$. As the result section will show, this turns out to be much more efficient in our task than the original purely exploratory agent proposed in [28]. This nevertheless does not outperform the meta-learning algorithm proposed in this paper. We finally tested the above-mentioned active exploration method proposed in computational neuroscience [29], [30]. The softmax function is also based on a weighted sum of Q-valuers and *exploration bonuses*. Nevertheless, the bonuses used are here computed as a low-pass filter on the square of δ computed by (1), which gives a simple approximation of the uncertainty associated to each Q-value.

Algorithm 1 Active Exploration With Meta-Learning

```
Initialize ω^A_{i,0}, ω^C_{i,0}, Q_{i,0}, β_0 and σ_0
for t = 0, 1, 2, ... do
    Select discrete action a_t with softmax(s_t, β_t) (Eq. 3)
    Select action parameters θ̃^a_{i,t} with GaussianExploration(s_t, a_t, θ^a_{i,t}, σ_t) (Eq. 4)
    Observe new state and reward {s_{t+1}, r_{t+1}} ← Transition(s_t, a_t, θ̃^a_{i,t})
    Update Q_{t+1}(s_t, a_t) in the discrete Q-Learning (Eq. 2)
    Update function approx. ω^C_{i,t+1} and ω^A_{i,t+1} in continuous actor-critic (Eq. 5, Eq. 6)
    if meta-learning then
        Update reward running averages r̄_t and r̿_t (Eq. 7)
        Update β_{t+1} and σ_{t+1} (Eq. 8)
    end if
end for
```

III. Numerical Experiments

A. Global Experimental Paradigm

The global experimental paradigm adopted here simulates a robot interacting with a human and trying to maximize the human engagement in the task by dynamically adjusting its behavior. We do not pretend to model all aspects of real human engagement. Instead, we simply simulated a partial measure of human engagement during interaction with a robot which has been suggested by [7]: this engagement represents the human's attention toward the robot and its actions, proposed to be estimated in real settings through measures of human gaze and body posture. The task consists in having the robot point toward one among a set of discrete objects (e.g., cubes on a table) while varying continuous parameters of action which here abstractly represent the expressivity of the action (i.e., for how long the robot moves its hand back and forth; with which angle the robot bends its torso) aimed at making the pointing gesture more explicit.

The paradigm is based on a currently ongoing pilot experiment, conducted in a specially arranged laboratory setting, where the NAO robot interacts with children with autistic spectrum disorders (presenting different levels of symptom severity, according to predefined assessment criteria), and tries to engage them in collaborative action by pointing at desired objects. In this pilot experiment, the human engagement processed by the robot as a reward is low when the child does not pay attention to the robot and its action, increases mildly when the child starts to look but remains far away, increases further when the child comes closer to look, and becomes maximal when the child helps the robot catch the object. At different moments in time, and playing with different objects, the robot explores different levels of expressivity until finding an appropriate level specific to each child it interacts with. Nevertheless, its behavioral exploration can sometimes either: 1) make the child engagement suddenly drop or 2) transiently low when the child's attention is captured away for a few seconds. The robot should thus adapt its action parameters in the first case while ignoring engagement perturbations in the second case.

The results of the pilot experiment with real children (12 children so far) are for the moment preliminary. More trials with more children are planned for the near future, to be conducted as an interventional study in a special education school, which will aim at more systematically evaluating the full potential of the approach. Nevertheless, while the analysis of the results of these studies is out of the scope of this paper and will rather be the focus of a future publication, preliminary results of the first pilot study are quite promising. More precisely, eight children successfully increased their engagement, although not optimally, ending up moving the pointed object closer to the robot, and moreover expressed in a post-interview that they found the task relatively easy and that they would like to play more often with the robot. In addition, the initial findings already highlight that there exists a large variance in the behavior and the preferences of children in such child–robot interaction scenarios (which are apparently not only related to cognitive age), providing evidence that an online active exploration process combined with RL is necessary for the robot to adapt to such variations of human engagement, which may consequently have a significant effect in terms of enhancing the targeted social responses of the child. To our knowledge, such an adaptive robot learning algorithm does not yet exist. We present here simulations and robustness analyses of this novel algorithm in order to propose a feasible solution to such a human engagement maximization problem during HRI.

B. Nonstationary Multiarmed Bandit

At first we evaluate the algorithm's performance on a nonstationary multiarmed bandit problem in order to estimate its intrinsic adaptive characteristics, as the single-state HRI which will be used in the next section can also be viewed as a nonstationary multi armed-bandit task with continuous parameterized actions. Here, we compare our meta-learning algorithm (modified and simplified accordingly to fit a single-state setup) with the performance of SW-UCB [27], D-UCB [37], and UCB1 [38]; the former two constitute analytically and experimentally proven algorithms on nonstationary cases.

Even though multiarmed bandits may seem to be out of the scope of our research, each state in an RL framework can be viewed as a multiarmed bandit problem, with the transition function defining the sequence on which each bandit is being visited according to the agent's actions. Our interest on such cases is crucial in order to better understand and improve its performance, as also to design an optimal decision-making agent in both high and low dimensional state spaces. One would argue however, that the most proper bandit setup for the task of our interest would be the nonstochastic (since the changes on reward distributions may depend on the robot's previous actions). However, here we consider the same stochastic setup used in [27] for comparison with two state-of-the-art adaptive algorithms. In stochastic setups there is no contextual information regarding the transitions of reward distributions and hence the performance of our algorithm on nonstationary cases would depend mainly on the intrinsic adaptive behavior of meta-learning.

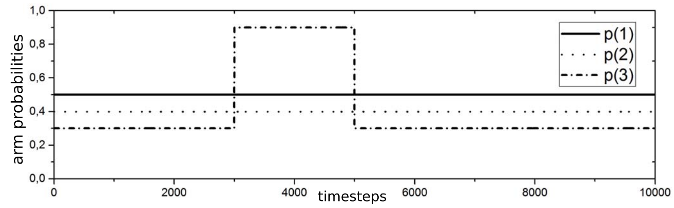


Fig. 1. Probability that an arm a will return a reward upon choice in the nonstationary multiarmed bandit task tested here. Adapted from [27].

In particular, the stochastic multiarmed Bernoulli bandit can be formulated as having a set of arms $\mathcal{K} = \{1, \dots, K\}$, each of them attached to a gambling machine, while at every episode $t \in \mathcal{T}$, with $\mathcal{T} = \{1, \dots, T\}$ denoting the sequence of decision episodes, the decision maker pulls an arm $a \in \mathcal{K}$ and receives a reward $r_t(a)$ with some unknown probability $p_t(a)$, and zero otherwise. For the *switching* task of [27], $K = 3$, $T = 10\,000$, the reward are binary $r_t \in \{0, 1\}$, $p_t(1) = 0.5$, $p_t(2) = 0.3$, $p_t(3) = 0.9$ for $3000 \leq t < 5000$, and $p_t(3) = 0.3$ otherwise as seen in Fig. 1.

For our implementation we used the Boltzmann softmax of (3) for action selection, while we updated Q-values according to a simple learning rule with a learning rate of 0.4. For updating the inverse temperature β of the softmax function, we used an iterative procedure with the use of a simplistic affine as defined in (8)

$$\beta_{t+1} \leftarrow \max\left\{0, \beta_t + \mu\tau_2\Delta\bar{\bar{r}}_t + \epsilon\right\} \tag{9}$$

using $\mu = 0.25$, $\tau_1 = 20$, $\tau_2 = 300$, and ϵ as a very small constant to ensure increment of exploitation on long stationary intervals. For hyper-parameter tuning we performed grid search on a large scale, observed areas of optimal behavior and robustness and rerun grid search on smaller regions until sufficient performance was achieved. All simulations were repeated for 500 sessions, and the average total regret per episode, the final cumulative regret and the final cumulative reward for each session were computed.

As seen in the results of Fig. 2, UCB-1 outperforms all other algorithms from episodes 1 to 3000 as expected, due to the stationary nature of this interval, but suffers from large regret values and learning inertia right after the first change point. SW-UCB is initially the second best, demonstrating a balanced exploration-exploitation ratio which is interpreted by the small average slope of the graph. The adaptive nature of meta-learning is exhibited right after the first change point. The flat line of the graph during timesteps 3500–5000 demonstrates that the action taken during this interval was the optimal, achieving an almost "no-regret" performance. At the end of episode 5000 SW-UCB has the lowest cumulative regret. However, the average slope of the graph is approximately equal with the one of the first interval, inferring the same levels of exploration regardless the large "probability gap" between the optimal and the second best action. After the second change point, the gap between the optimal and the second best action is once again small, and SW-UCB performs better than all others except for the last 1500 timesteps where UCB-1 has overcome the learning inertia. Yet UCB-1

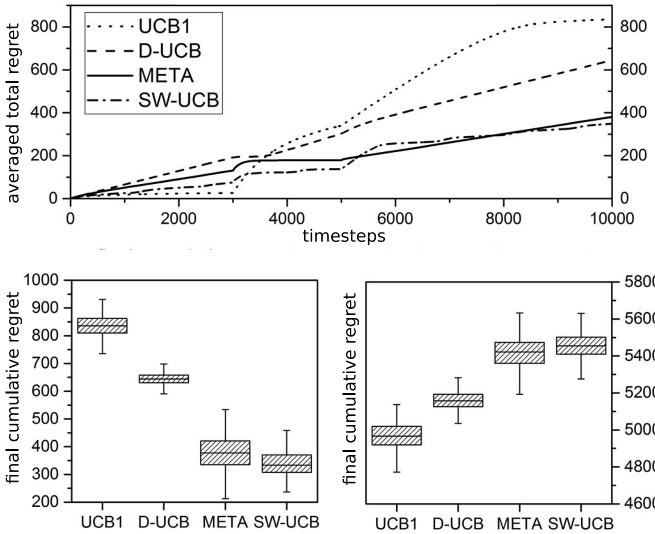


Fig. 2. Performance in the tested abruptly changing bandit task shows that the proposed meta-learning algorithm performs as well as SW-UCB. Top: The averaged cumulative regret per episode. Bottom left: The final cumulative regret for 500 sessions. Bottom right: The final cumulative reward for 500 sessions.

has already accumulated large regret. Finally, the overall performance of meta-learning algorithm is comparable with the one of SW-UCB, despite its multistate nature.

Even though for the proposed problem set SW-UCB and D-UCB used as parameters the ones that guarantee an upper bound of regret as shown in [27], meta-learning for bandits was empirically optimized through an extensive parameter grid search. In [39], however, new evidence about the empirical performance of all the above is provided, with meta-learning (MLB algorithm) achieving significantly better performance in most cases, while it can be also enhanced with the use of sibling Kalman filters. More precisely, [39] thoroughly studied cases with altering volatility levels of the environment, as well as different probability gaps between the optimal and the second best actions, demonstrating the intrinsic adaptive nature of our algorithm at the lower level of an RL framework.

C. Simple HRI Simulation

We then test the algorithm described in Section II in a simple simulated HRI task involving a single state and six discrete actions (corresponding to pointing toward different cubes on a table), hence in essence similar to the nonstationary multi-armed bandit paradigm. However, a major difference here is the requirement for the robot to not only learn to perform the optimal discrete action $a^\star$ (i.e., pointing at the cube that the human is interested in) but also to perform it with the optimal continuous parameters of action $\mu^\star$ ($\mu^\star \in [-100; 100]$). These continuous parameters of action abstractly represent different properties of movement, such as velocity, direction, strength, expressivity, or any other aspect which could affect the human engagement in the task. In other words, rather than associating a fixed probability of reward to each discrete action, an action will yield reward only when its continuous parameters are chosen within a Gaussian distribution around the current optimal action parameter $\mu^\star$ with variance $\sigma^\star$ (which are unknown to the robot). This mimics the fact that, depending on the human the robot is interacting with, its action should neither be executed too fast nor too slow, should neither be too expressive nor too little expressive. For each interlocutor, there are appropriate continuous parameters of action that the robot needs to find autonomously. Finally, every n timesteps, $a^\star$ and $\mu^\star$ change—representing a change in the robot behavior that maximizes the engagement of the simulated human—so that the task is nonstationary and requires constant re-exploration and learning by the robot. In Sections III-C and III-D, these abrupt task changes mimic the case where the human at some point changes its object of interest and wants the robot to also change its way of interacting with this object (e.g., faster). In Section III-E, the object of interest of the human does not change (same cube) but the abrupt task change corresponds to a transient perturbation of the human engagement (e.g., the human's attention is attracted away by the noise of someone else entering the room) that the robot has to robustly cope with in order not to deviate from the task at hand.

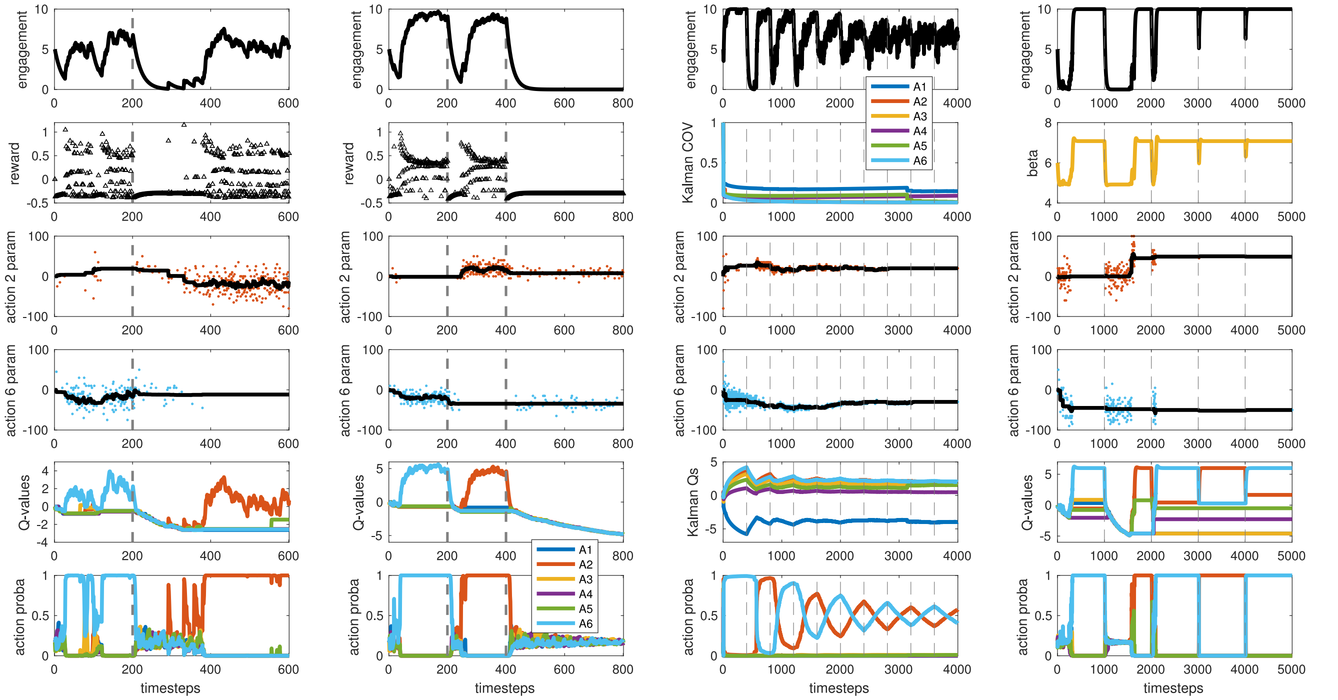
Previous researches on HRI have shown that the human engagement can be a critical aspect of the quality of the interaction [7]. Nevertheless, during interaction tasks the actions performed by a robot can have delayed effects on the human's behavior and on his engagement. To mimic this, we chose the reward to be given by a dynamical system which is based on the virtual engagement $e(t)$ of the human in the task. This engagement somehow represents the attention that the human pays to the robot and will constitute a reward signal, since this type of joint attention social signals have been shown to activate the same brain regions that are activated by nonsocial extrinsic reward such as food or money [40]. The simulated human engagement $e(t)$ starts at 5, increases up to a maximum $e_{\max} = 10$ when the robot performs the appropriate actions with the appropriate parameters, and decreases down to a minimum $e_{\min} = 0$ otherwise

$$e_{t+1} = \begin{cases} e_t + \eta_1(e_{\max} - e_t)H(\theta_t^a), & \text{if } a_t = a^\star \text{ \& } H(\theta_t^a) \geq 0 \\ e_t - \eta_2(e_{\min} - e_t)H(\theta_t^a), & \text{if } a_t = a^\star \text{ \& } H(\theta_t^a) < 0 \\ e_t + \eta_2(e_{\min} - e_t), & \text{otherwise} \end{cases} \quad (10)$$

where $\eta_1 = 0.1$ is the increasing rate, $\eta_2 = 0.05$ is the decreasing rate, and $\mathcal{H}(x)$ is the re-engagement function given by $\mathcal{H}(x) = 2(\exp(-[(x - \mu^\star)^2]/2\sigma^{\star 2}) - 0.5)$ where $a^\star$, $\mu^\star$, and $\sigma^\star$ are, respectively, the optimal action, action parameter, and variance around $a^\star$.

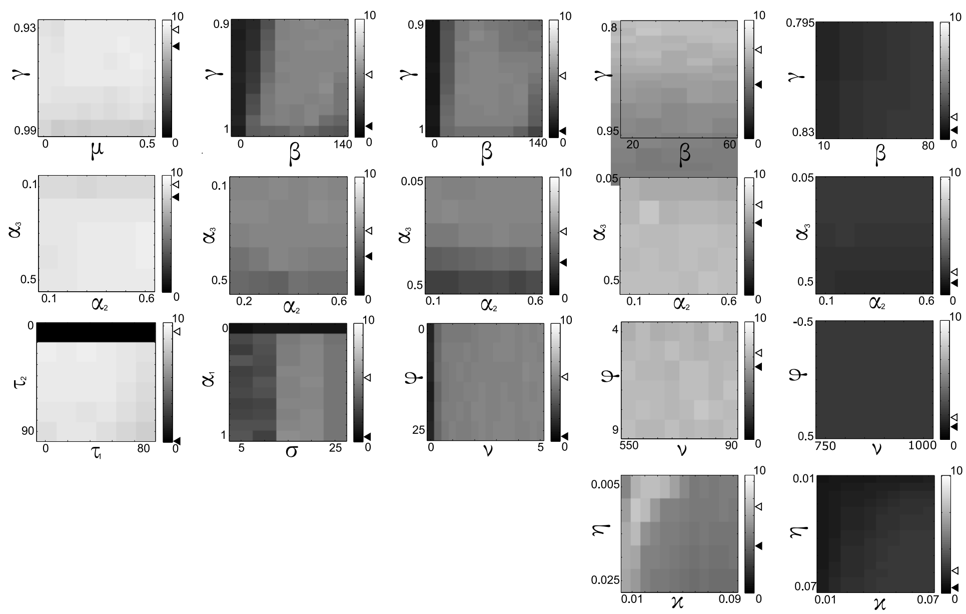
The reward function is then computed as $r(t+1) = e(t+1) + \lambda\Delta e(t)$ where $\lambda = 0.7$ is a weight and $\Delta e(t) = e(t+1) - e(t)$. This reward function ensures that the algorithm gets rewarded in cases where the engagement $e(t+1)$ is low but nevertheless has just been increased by the action tuple $(a(t), \theta^a(t))$ performed by the robot.

We first present a set of short simulations of different ways to handle exploration in the algorithm (shown in Fig. 3) in slightly different task conditions just to illustrate the strengths and weaknesses of the tested alternative solutions. We will later show proper comparisons of these methods in the exact same task conditions in order to assess their performance. We



(a) (b) (c) (d)

Fig. 3. Simulations of the parameterized RL algorithm with different methods to handle exploration. (a) Fixed $\sigma = 20$ and $\beta = 4$ (no active exploration) can adapt to abrupt task changes but does not maximize simulated human engagement. (b) Fixed $\sigma = 10$ and $\beta = 4$ (no active exploration) can maximize engagement and adapt to some task changes but not in the case where the new optimal action parameter is too far away from the previous one. (c) σ_t^a and b_t^a tuned by Kalman-RL (active exploration) can adapt to multiple consecutive task changes but will overall progressively average the statistics of the different task conditions. (d) σ_t and β_t tuned by meta-learning (active exploration) can maximize engagement and adapt to task changes after fast transient re-exploration phases. Gray vertical dashed lines indicate changes in optimal action tuple.



(a) (b) (c) (d) (e)

Fig. 4. Parameter optimization for the different tested algorithms. For each algorithm, each datapoint corresponds to the average engagement obtained for ten simulations of the task with a given parameter set. For each given model and pair of parameters, black full and empty arrow heads on the colorbar, respectively, indicate the maximal and minimal mean engagement reached within the subplot. (a) Active exploration with meta-learning. (b) Passive exploration. (c) Active exploration with exploration bonuses. (d) Active exploration with Kalman-QL. (e) Fully exploratory Kalman-QL.

first simulated the algorithm without active exploration (thus with a fixed $\sigma = 20$) in a task, where the optimal action tuple $(a^\star, \mu^\star)$ is $(a_6, -20)$ during 200 timesteps ($\sigma^\star = 10$ in all the experiments presented here), then switches to $(a_2, -20)$ until timestep 600. Fig. 3(a) shows that the algorithm first learns the appropriate action tuple $(a_6, -20)$, then takes some time to learn the second tuple, making the engagement drop between timesteps 200 and 400 and eventually finds the second optimal tuple. Nevertheless, $\sigma = 20$ makes the robot select action parameters $\widetilde{\theta}_t^a$ with a large variance (illustrated by the clouds of dots around the learned action parameters θ_t^2 and θ_t^6 plotted as black curves). As a consequence, the engagement is not optimized and always remains below 7.5. In contrast, the same algorithm with a smaller fixed variance $\sigma = 10$ can make the engagement reach the optimum of 10 when the optimal action tuple is learned [Fig. 3(b) before timestep 400], but results in too little exploration which prevents the robot from finding a new action parameter when it is too far away from the previously learned one (after timestep 400, the new optimal action tuple is $(a_6, 20)$). These two examples illustrate the need to actively vary the variance σ_t as a function of changes in the robot's performance.

We next tested active exploration with *exploration bonuses* based on the Kalman-RL algorithm [28] in a task alternating between optimal tuples $(a_2, -20)$ and $(a_6, 20)$ every 400 timesteps. Fig. 3(c) shows the results of this extended version of Kalman-QL. The diagonal terms of the covariance matrix COV in the Kalman filter nearly monotonically decrease, resulting in a large variance σ_t when action a_6 is executed until about timestep 600, and progressively decreasing the variance until the end of the experiment. Nevertheless, the algorithm quickly finds the appropriate action parameters and rapidly shifts between actions a_2 and a_6 after each change in the task condition. In the long run, the model progressively averages the statistics of the two conditions and learns to perform both actions with 50/50 probabilities [bottom part of Fig. 3(c)] which decreases the simulated engagement (top).

We then tested active exploration with the meta-learning algorithm in a slightly more difficult task where the optimal action tuple $(a^\star, \mu^\star)$ alternates between $(a_2, -50)$ and $(a_6, 50)$ every 1000 timesteps [Fig. 3(d)]. Transient drops in the engagement result in transient decreases in the exploration parameter β_t as well as transient increases in the variance σ_t. This enables the algorithm to go through quick transient but wide exploration phases and to rapidly reconverge to exploitation, thus maximizing the simulated engagement. Strikingly, the engagement decreases less and less after each change in task condition (i.e., timesteps 1000, 2000, 3000, and 4000), which shows that the algorithm adapts faster and faster to task changes. Note that this simulated engagement is indicative of the robot's behavioral accuracy because it increases according to (10) only when the robot performs the optimal discrete action $a^\star$ with a continuous parameter θ_t^a close to the optimal parameter $\mu^\star$. Thus, engagement and behavioral accuracy are correlated here, and when the simulated engagement reaches 10, this means that the robot performs the optimal behavior 100% of the time thanks to the increase of β_t and decrease of σ_t according to (8) which focuses the algorithm on pure exploitation once the optimal behavior is reached.

We performed an exhaustive search of the parameters that permit each algorithm to reach its highest performance in the difficult version of the task (Fig. 4). While the previous simulations used different task conditions to illustrate the respective properties of each tested algorithm, here, all algorithms are thus compared on the same task in order to compare their performance. Active exploration based on meta-learning reached the highest performance, with an average engagement of 9.2 obtained with the best parameter-set. Importantly, the performance was robust for a large portion of the explored parameter space, except in the case where $\tau_2 = 1$ for which the mean simulated engagement during the experiment was 0.02. In all other cases (we tested all combinations of $(\tau_1, \tau_2) \in \{1, 2, 5, 10, 50, 90\}^2$, thus including cases where $\tau_1 = \tau_2$, cases where $\tau_1 < \tau_2$ and cases where $\tau_1 > \tau_2$), the mean simulated engagement is higher or equal to 8.09, thus higher than the best engagement obtained for all other tested algorithms (Fig. 4). Interestingly, the original meta-learning paper using these two parameters for active exploration [26] (which does nevertheless neither include the continuous parameters of actions nor the Gaussian-exploration process proposed here) only presented simulations were $\tau_1 = \tau_2$, thus leaving the question open whether different values would also work or not. Thus, the exploration of the parameter space presented here shows that the choice of τ_1 and τ_2 for the tasks studied here is not crucial.

Active exploration based on Kalman-QL gave the second best performance, with an average engagement of 7.2. Interestingly, the original fully exploratory Kalman-QL agent proposed in [28] did not manage to get an average engagement higher than 3, due to the nonstationarity of the environment. Similarly, the tested computational neuroscience method for the estimation of *exploration bonuses* did not reach an average engagement higher than the passive exploration algorithm. This is due to the presence of the engagement variation term in the reward function, which makes the reward decrease once a plateau of engagement is reached, leading to non-null reward prediction errors and thus to non-null exploration bonuses when the algorithm should rather be exploiting. Finally, Fig. 5 shows the average and standard deviation of the simulated engagement obtained for these ten simulations of the task with the two best algorithms and the passive exploration one. The blue curve shows the performance of the algorithm without active exploration (i.e., fixed $\sigma = 19$ obtained through parameter optimization), which adapts to each new condition but never exceeds a plateau of about 6. The green curve shows the active exploration with Kalman, which adapts faster at the beginning but progressively decreases its maximal engagement. The red curve shows the active exploration with meta-learning which initially takes more time to adapt but then only performs short transient explorations and reaches the optimum engagement of 10.

D. Realistic HRI Simulation

In order to have a more realistic demonstration of the proposed algorithm and to gain a better insight of its envisaged application to HRI tasks, we created and visualized a scenario using the V-REP robot simulator (Fig. 6). In the considered scenario, a small humanoid robot, in this case a NAO, interacts with a human subject, where the envisaged goal is to collaboratively perform a task involving pointing at, picking

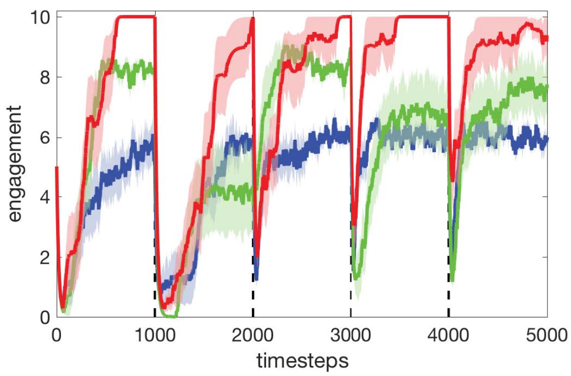


Fig. 5. Comparison of engagement in ten simulations of the meta-learning model (red), the model without active exploration (blue), and the Kalman-QL (green).

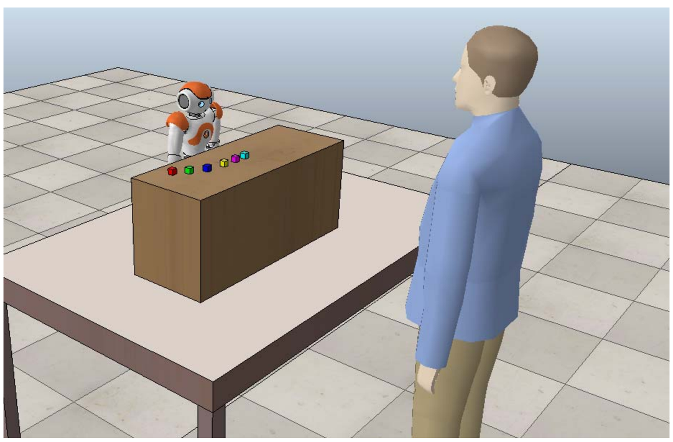


Fig. 6. Snapshot of the V-REP simulated HRI scenario showing the configuration of the experiment.

up, and placing objects in the scene in order to build a puzzle. Such a collaborative HRI scenario is in line with the objectives defined in the frames of the EU-funded project BabyRobot (H2020-ICT-24-2015-6878310), where a set of child–robot interaction use-cases have been designed and are currently being implemented to study the development of specific socio-affective, communication, and collaboration skills in children. In particular, the task considered in the simple scenario simulated here comprises a set of 6 objects (cubes) set in front of the human and the robot (Fig. 6). Each robot action at the current implementation stage corresponds to a *pointing gesture* of the robot (with its right arm) toward one of the 6 cubes. The human engagement is expressed through the gazing direction with respect to the pointed cube. In essence this means that, if the engagement is high, the attention of the human subject is directed toward the pointed cube, while if the engagement is low, the human turns his head around.

In the current implementation, the human gaze is sampled from a normal distribution centered around the position of the object corresponding to the action (pointing gesture) currently performed by the robot, with a standard deviation that depends (inversely) on the current human engagement value. Changes of human gaze direction are sampled and executed every T_1 time-steps, while each robot action is performed and remains unchanged for T_2 time-steps ($T_2 = nT_1$); meaning that the robot is assumed to collect n observations of human gaze direction changes before selecting and executing a new action. In this section, a simple case is considered where the actual simulated human engagement value is assumed to be directly known to the robot, which nevertheless poses the problem of being able to quickly adapt to engagement changes. The next section will then show tests of a more realistic scenarios where the actual human engagement value will be assumed to be unknown to the robot and estimated online based on observations of the human gaze directions. Furthermore, the action parameter will also be integrated in the task and will represent a measure of the overall "intensity" of the robot's arm movements when executing a communicative action (e.g., a pointing gesture).

It is interesting in this scenario to study and visualize the performance of the proposed meta-learning active exploration algorithm when the optimal action parameter changes (while the optimal action itself remains the same). Fig. 7 compares the performance of the proposed meta-learning algorithm with the Kalman Q-learning mechanism, when the optimal action parameter undergoes a 50% change (from a value of -50 to -25). We can see that the meta-learning algorithm adapts much faster to the new task parameter. Specifically, the human engagement drops to no less than 70% of the maximum engagement and recovers to 85% after a few trials (in this case, after approximately 25 trials). In addition, the action parameter converges fast to the optimal value (in this example, after 30 trials). On the contrary, the Kalman Q-learning algorithm fails to adapt to the new task parameter and to raise the engagement back to its maximum value, resulting in a suboptimal engagement for the rest of the experiment. This behavior is also illustrated by the oscillation of the action parameter as it fails to converge to the optimal value.

These initial simulations provide a first understanding of practical considerations that will have to be addressed toward the implementation and deployment of more realistic HRI scenarios as already described. Initial results are promising showing the potential of the proposed meta-learning algorithm as a scheme to efficiently adapt to nonstationary conditions in challenging HRI scenarios.

E. Engagement Estimation Process in the HRI Simulation

We finally made a last experiment where we consider that the human engagement is unknown, can be the subject to transient and more-or-less long-lasting perturbations (e.g., the human's attention is attracted away by the noise of someone else entering the room), and that the robot has to estimate this engagement online based on nonverbal cues expressed by the human partner during HRIs. This experiment is aimed at making the simulated HRI scenario even more realistic and obtain a more reliable assessment on the applicability of the developed learning algorithms in real use-case scenarios where the human subject can be disturbed during the task, and the robot should avoid unlearning the correct behavior because of this perturbation. Thus, in this last experiment we focus on a task where there is a single object to focus

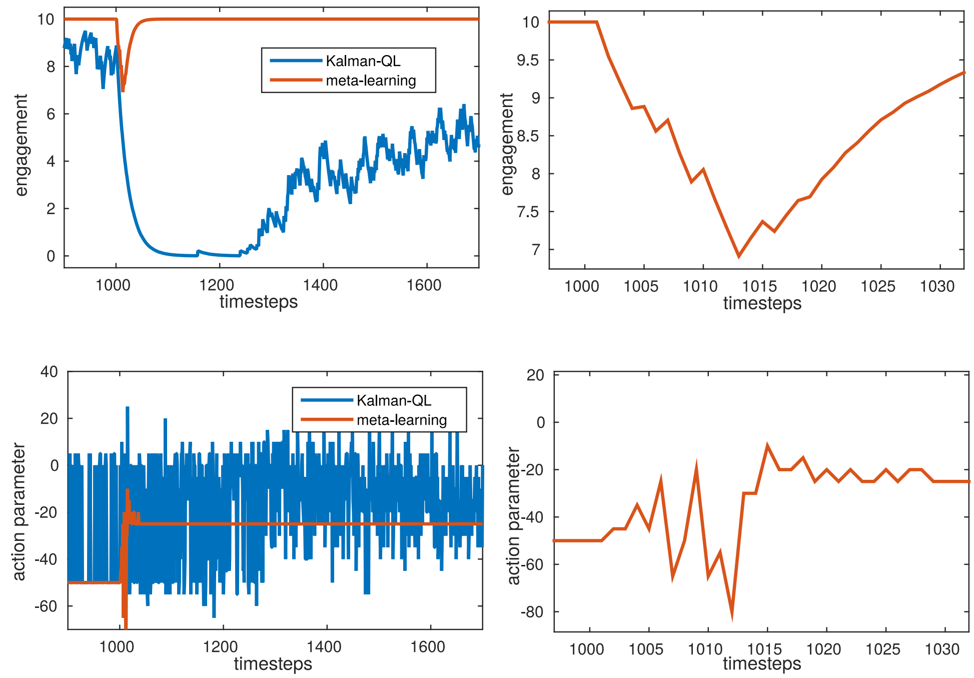


Fig. 7. Left column: Comparison of engagement (top) and action parameter (bottom) convergence between the meta-learning and the Kalman-QL algorithms. Right column: Zoomed-in plots depicting convergence performance of the proposed meta-learning algorithm.

on (but the robot still has to adapt its continuous parameters based on variations of human engagement), and we conduct an initial evaluation as to how scalable and generalizable the proposed learning algorithm is in a more close to real-life scenario and how the presence of an uncertainty on human engagement estimation may affect the performance of the system (Fig. 8).

As mentioned in [7] and [41], head pose data is proved to be highly correlated with human's engagement. In particular, clustering of the gaze, head stability as well as head pose and its variance constitute important features for the evaluation of human engagement in face-to-face, interactive scenarios. In the current implementation of our simulations, the human head pose changes according to his engagement. More specifically, pitch and yaw angles of the head are each generated by sampling a normal distribution centered around the position of the object pointed by the robot. The distribution's standard deviation is inversely proportional to human engagement. Thus, when the engagement drops, the head pose variance increases, meaning that the human is disengaged from the task and starts looking around. On the contrary, when engagement is high the attention of the person is focused on the pointed object which results in a stable head with low variance. This dependence is illustrated in Fig. 8.

The engagement estimation is achieved by measuring the mean standard deviation (MSD) of the human's head pitch and yaw angles with respect to the cube's location projected on the pitch-yaw plane pointed by the robot in a specified time window. In particular, the robot collects n observations of human's head pose before selecting and executing a new action. Given that the head pose is measured by visual means, the measurement error is taken into account and modeled as an additive

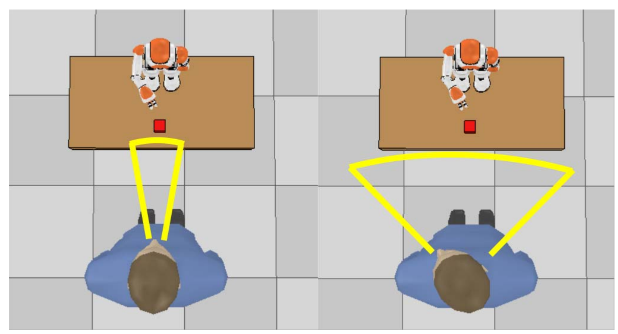


Fig. 8. Left: Low head pose variance. Human engagement is high. Right: High head pose variance. Human engagement is low.

Gaussian noise with zero mean and standard deviation σ that depends on the accuracy of the visual head pose estimation.

The presence of uncertainties in observed measurements (head pose) for the estimation of an unknown variable (engagement) pushed us to use a Kalman filter for producing more accurate engagement estimates. Since the engagement model is unknown to the robot, the prediction step of the Kalman filtering process considers an engagement estimate based on the head pose variance. Specifically, we consider a re-engagement function, similar to $\mathcal{H}(x)$ that was defined earlier, given by

$$\hat{\mathcal{H}}(s_p, s_y) = 2\big(\exp\big(-k_1 \cdot s_p - k_2 \cdot s_y\big) - 0.5\big) \tag{11}$$

where s_p and s_y are the head pitch and yaw MSD from the pointed cube in a time window and k_1 and k_2 are positive constants. It is clear that $\hat{\mathcal{H}}$ is maximized when the head pose variance is zero, or when human engagement is maximum. Therefore, the estimator increases the estimated engagement up to 10 when the human's head pose variance is

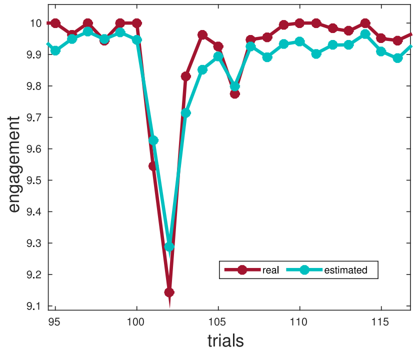


Fig. 9. Real (simulated) versus estimated engagement based on $n = 5$ head pose observations per trial and Gaussian measurement noise with $\sigma = 0.5$.

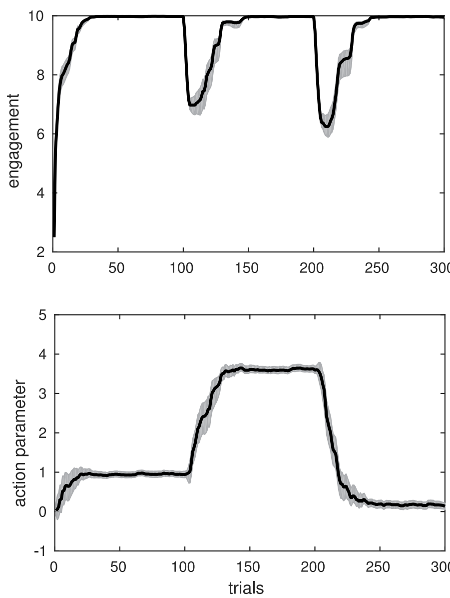


Fig. 10. Top: Real (simulated) human engagement (mean and variance after 20 runs) when the optimal action parameter undergoes step changes every 100 trials as shown in bottom figure (engagement does not drop below a 60% value). Bottom: Executed action parameter (mean and variance after 20 runs) involving step changes.

low ($\hat{\mathcal{H}}(s_p, s_y) > 0$) and decreases it down to 0 otherwise ($\hat{\mathcal{H}}(s_p, s_y) < 0$). Of course this is a rough and unreliable estimate of the engagement, since it is based on noisy pitch and yaw measurements. However, accuracy improves in the update step, where the measurement noise is taken into account and the estimate is updated according to the optimal Kalman gain. Fig. 9 shows the accuracy of the engagement estimation when the robot collects five head pose observations per trial and the measurement noise has $\sigma = 0.5$. After the human engagement is evaluated through the described process, it is provided to the robot as a reward. The reward function now considers the estimated engagement $\hat{e}$ and is computed as $r(t+1) = \hat{e}(t+1) + \lambda \Delta \hat{e}(t)$.

The first series of numerical experiments that we conducted involves step changes in the optimal (continuous) action parameter performed every 100 trials. A series of 20 runs for the same step-change scenario has been conducted and the results are shown in Fig. 10. In this situation the optimal action parameter increases from 1 to 4 (300% increase) and after 100 trials drops to 0. We calculated the mean value and standard deviation of the actual executed action parameter and the human engagement at every timestep. The results indicate that although the optimal action parameter initially quadrupled, the adaptation was fast enough to keep the engagement above a 70% value and consistently make it converge to a value above 90% after approximately 25 trials. In the next 100 trials the action parameter change was larger (dropped from 4 to 0) leading to a slightly wider engagement drop.

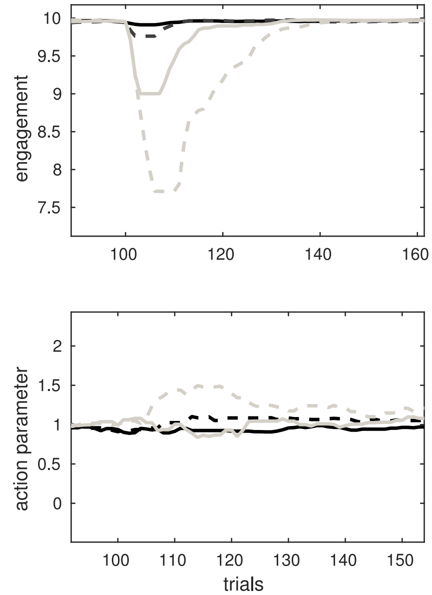


Fig. 11. Human engagement (top) and action parameter (bottom) during optimal action parameter perturbations with duration of 1 (solid black), 2 (dashed black), 5 (solid gray), and 10 (dashed gray) trials, respectively.

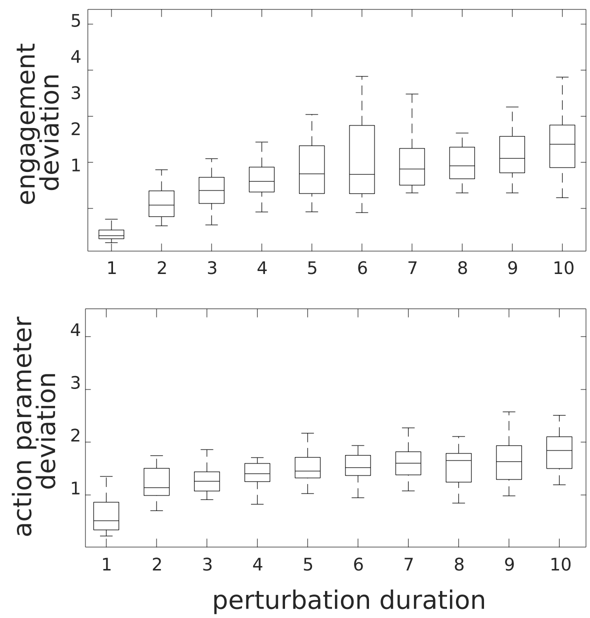


Fig. 12. Performance for increasing perturbation duration (number of trials). Top: Engagement deviation from its maximum value (10). Bottom: Action parameter deviation from its optimal value (1). The measurement noise has $\sigma = 1$.

TABLE I
NUMBER OF TRIALS NEEDED FOR ENGAGEMENT TO REACH 90% OF ITS MAXIMAL VALUE AFTER A PERTURBATION

Perturbation duration	Mean	STD	25% percentile	75% percentile
1	1.8	1.5	0	3
2	3.95	1.5	3	4.5
3	8.85	7.6	3.5	11
4	11.6	9.5	4	15.5
5	11.05	6.1	7	11.5
6	11.5	5.1	8	15
7	12.8	4.5	9.5	16
8	15.15	6.4	11	18
9	14.35	5.5	11	18
10	16.45	6.6	12	19

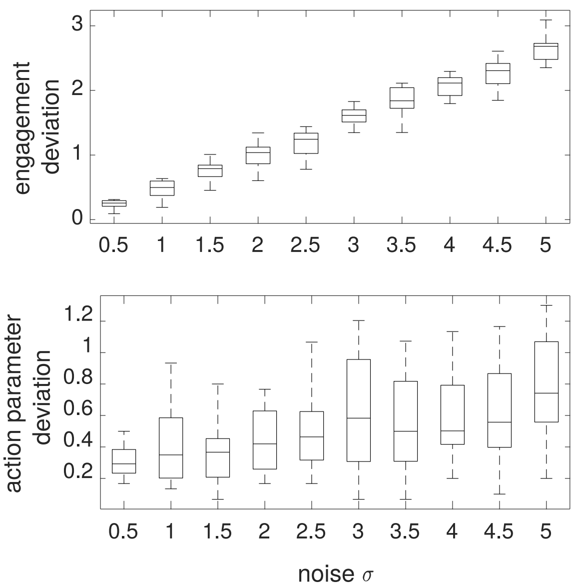


Fig. 13. Performance for increasing noise σ. Top: Engagement deviation from its maximum value (10). Bottom: Action parameter deviation from its optimal value (1).

During a real HRI task, it is natural for a person and much more for a child to be distracted by an external event (loud noise, presence of other people, etc.). We simulate such a perturbation as an abrupt and short in time (impulse-type) change of the optimal action parameter. The behavior of the algorithm is depicted in Fig. 11 for various durations of the perturbation impulse that begin at trial 100. In these experiments the optimal action parameter has a value of 1 that is changed to 5 during the perturbations. As the perturbation duration increases, the engagement and action parameter deviation from their optimal values become larger. However, we observe that when the perturbation is short (1–2 trials), the executed action parameter is almost unaffected and the human engagement drop is unnoticeable. In order to further quantify the performance of the algorithm, we calculate the mean absolute deviation (MAD) of the real (simulated) engagement and the action parameter from their optimal values for perturbation duration in the range of 1–10 trials. Fig. 12 indicates that longer perturbations lead to slower adaptation, resulting in larger MAD values. The same results are also numerically shown in Table I that presents the number of trials needed for the engagement to recover 90% of its maximal value after the end of the perturbation. It should also be highlighted, though, that as illustrated by the obtained results, no matter how long the perturbation, the algorithm will always reconverge to the optimal value.

In a similar way, Fig. 13 shows the engagement and action parameter deviation for an increasing σ of the Gaussian head pose measurement noise. In the particular experiment the perturbation duration is 1 trial. It is clear that noisier pitch and yaw head angles measurements result in larger deviations of the estimated engagement from the real (simulated) engagement (Fig. 13 top). The same holds for the action parameter whose deviation from its optimal value is proportional to the amplitude of the measurement noise (Fig. 13 bottom).

IV. CONCLUSION

In this paper, we have shown that a meta-learning algorithm based on online variations of reward running averages can be used to adaptively tune two exploration parameters simultaneously used to select between both discrete actions and continuous action parameters in a parameterized action space.

We first compared the proposed algorithm with standard bandit methods in the nonstationary (switching) multiarmed bandit task proposed by [27]. We showed that it reaches a performance which is not different from one of the state-of-the-art bandit methods, namely SW-UCB. Interestingly, SW-UCB does not adapt well to some other nonstationary tasks [42]. Moreover, bandit methods work specifically in single-state tasks. The meta-learning algorithm proposed here seems promising in that it is in principle generalized to continuous actions and multistep tasks. In future work, we will compare it with bandit methods in a variety of nonstationary tasks and then study its performance in sequential multistep tasks.

We then applied the proposed meta-learning algorithm to a simple simulated HRI task consisting in having the robot point toward one among a set of discrete objects (e.g., cubes on a table) while varying continuous parameters of action which here abstractly represent the expressivity of the action (i.e., for how long the robot moves its hand back and forth; with which angle the robot bends its torso) aimed at making the pointing gesture more explicit. The task involved abrupt task changes mimicking either the case where the human at some point changes its object of interest and wants the robot to also change its way of interacting with this object (e.g., faster), or the case where a transient perturbation of the human engagement (e.g., the human's attention is attracted away by the noise of someone else entering the room) requires the robot to show robustness in order not to deviate from the task at hand.

Previous studies have investigated ways to handle nonstationary, noisy, and delayed feedback during HRI [43], [44], especially engagement signals [14], [45]. Nevertheless, one of the novelties of this paper was the use of human engagement monitoring signals as a reward signal for robot RL during social interaction. Here, the proposed reward function consisted in a weighted sum of the human's current engagement and variations of this engagement (so that a low but increasing engagement is rewarding). We found that the active exploration meta-learning algorithm outperforms continuous parameterized RL both without active exploration and with active exploration based on alternative methods, such as uncertainty variations measured by a Kalman-Q-learning algorithm. While we had previously successfully used the Kalman Q-learning proposed by [28] to coordinate model-based and model-free RL in a stationary task [46], it was not appropriate for the current nonstationary task.

The robustness of the algorithm was then tested in situations where the human is distracted by external events and we showed that no matter the length of the perturbation, the algorithm would always come back to optimal behavior afterward. In fact, the algorithm succeeded to keep human engagement high when engagement perturbations were short. Then, we showed how engagement estimation is affected by the presence of measurement noise. Although the algorithm is not significantly affected by small noise amplitudes, the performance drops when uncertainties in human engagement are high as shown by the increased action parameter deviation from its optimal value. To improve this, the robot could reset its engagement estimation when the human looks at a discrete object whose location is known to the robot. The robot could even ask the human to look at the object in order to recalibrate its estimation. In future work, we will address these issues and test the algorithm in more complex simulated interaction tasks before applying it to real HRI.

The different results presented in this paper suggest that the proposed active exploration scheme in combination with the described engagement estimation process could be a promising solution for applications related to HRI tasks in dynamic environments.

Acknowledgment

The authors would like to thank K. Doya, B. Girard, O. Pietquin, B. Piot, I. Rano, O. Sigaud, and G. Viejo for useful discussions.

References

[1] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 65–78, Feb. 2002.

[2] T. Brick and M. Scheutz, "Incremental natural language processing for HRI," in *Proc. 2nd ACM/IEEE Int. Conf. Human–Robot Interact. (HRI)*, Arlington, VA, USA, 2007, pp. 263–270.

[3] J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson, "Towards human-like spoken dialogue systems," *Speech Commun.*, vol. 50, nos. 8–9, pp. 630–645, 2008.

[4] R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu, "Robust spoken instruction understanding for HRI," in *Proc. 5th ACM/IEEE Int. Conf. Human–Robot Interact.*, Osaka, Japan, 2010, pp. 275–282.

[5] S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier, "An architecture for fluid real-time conversational agents: Integrating incremental output generation and input processing," *J. Multimodal User Interfaces*, vol. 8, no. 1, pp. 97–108, 2014.

[6] S. Lemaignan, M. Warnier, E. Sisbot, A. Clodic, and R. Alami, "Artificial cognition for social human–robot interaction: An implementation," *Artif. Intell.*, vol. 247, pp. 45–69, Jun. 2017.

[7] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *Int. J. Soc. Robot.*, vol. 7, no. 4, pp. 465–478, 2015.

[8] M. Tomasello, *Origins of Human Communication*. London, U.K.: MIT Press, 2010.

[9] J.-D. Boucher *et al.*, "I reach faster when i see you look: Gaze effects in human–human and human–robot face-to-face cooperation," *Front. Neurorobot.*, vol. 6, p. 3, May 2012.

[10] S. Al Moubayed, G. Skantze, and J. Beskow, "The FURHAT back-projected humanoid head–lip reading, gaze and multi-party interaction," *Int. J. Humanoid Robot.*, vol. 10, no. 1, 2013, Art. no. 1350005.

[11] H. Admoni, A. Dragan, S. S. Srinivasa, and B. Scassellati, "Deliberate delays during robot-to-human handovers improve compliance with gaze communication," in *Proc. ACM/IEEE Int. Conf. Human–Robot Interact.*, Bielefeld, Germany, 2014, pp. 49–56.

[12] S. Bampatzia, V. Vouloutsi, K. Grechuta, S. Lallée, and P. F. M. J. Verschure, "Effects of gaze synchronization in human–robot interaction," in *Proc. Conf. Biomimetic Biohybrid Syst.*, 2014, pp. 370–373.

[13] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human–robot interaction," in *Proc. 5th ACM/IEEE Int. Conf. Human–Robot Interact. (HRI)*, Osaka, Japan, 2010, pp. 375–382.

[14] S. Ivaldi *et al.*, "Towards engagement models that consider individual factors in HRI: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task," *Int. J. Soc. Robot.*, vol. 9, no. 1, pp. 63–86, 2017.

[15] T. Ahmed and A. Srivastava, "A prototype model to predict human interest: Data based design to combine humans and machines," *IEEE Trans. Emerg. Topics Comput.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7885058/

[16] H. L. O'Brien and E. G. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," *J. Assoc. Inf. Sci. Technol.*, vol. 59, no. 6, pp. 938–955, 2008.

[17] M. Khamassi, S. Lallée, P. Enel, E. Procyk, and P. F. Dominey, "Robot cognitive control with a neurophysiologically inspired reinforcement learning model," *Front. Neurorobot.*, vol. 5, p. 1, Jul. 2011.

[18] W. Masson and G. Konidaris, "Reinforcement learning with parameterized actions," in *Proc. 13th AAAI Conf. Artif. Intell. (AAAI)*, Phoenix, AZ, USA, 2016, pp. 1934–1940.

[19] M. Hausknecht and P. Stone, "Deep reinforcement learning in parameterized action space," *arXiv:1511.04143*, 2015.

[20] J. Schmidhuber, "Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts," *Connection Sci.*, vol. 18, no. 2, pp. 173–187, 2006.

[21] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robot. Auton. Syst.*, vol. 61, no. 1, pp. 49–73, 2013.

[22] C. Moulin-Frier and P.-Y. Oudeyer, "Exploration strategies in developmental robotics: A unified probabilistic framework," in *Proc. IEEE 3rd Joint Int. Conf. Develop. Learn. Epigenetic Robot. (ICDL)*, Osaka, Japan, 2013, pp. 1–6.

[23] F. C. Y. Benureau and P.-Y. Oudeyer, "Behavioral diversity generation in autonomous exploration through reuse of past experience," *Front. Robot. AI*, vol. 8, pp. 44–93, Mar. 2016.

[24] J. X. Wang *et al.*, "Prefrontal cortex as a meta-reinforcement learning system," *Nat. Neurosci.*, vol. 21, no. 6, pp. 860–868, 2018.

[25] H. van Hasselt and M. A. Wiering, "Reinforcement learning in continuous action spaces," in *Proc. IEEE Symp. Approx. Dyn. Program. Reinforcement Learn.*, Honolulu, HI, USA, 2007, pp. 272–279.

[26] N. Schweighofer and K. Doya, "Meta-learning in reinforcement learning," *Neural Netw.*, vol. 16, no. 1, pp. 5–9, 2003.

[27] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," *arXiv preprint arXiv:0805.3415*, 2008.

[28] M. Geist and O. Pietquin, "Kalman temporal differences," *J. Artif. Intell. Res.*, vol. 39, no. 1, pp. 483–532, 2010.

[29] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for exploratory decisions in humans," *Nature*, vol. 441, no. 7095, pp. 876–879, 2006.

[30] M. J. Frank, B. B. Doll, J. Oas-Terpstra, and F. Moreno, "Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation," *Nat. Neurosci.*, vol. 12, no. 8, pp. 1062–1068, 2009.

[31] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas, "Active exploration and parameterized reinforcement learning applied to a simulated human–robot interaction task," in *Proc. IEEE Robot. Comput. Conf.*, Taichung, Taiwan, 2017, pp. 28–35.

[32] T. Tsitsimis, G. Velentzas, M. Khamassi, and C. Tzafestas, "Online adaptation to perturbations in human engagement during human–robot interaction with parallel reinforcement learning processes," in *Proc. MultiLearn Workshop 25th Eur. Signal Process. Conf.*, 2017, pp. 1–6.

[33] C. J. C. H. Watkins and P. Dayan, "*Q*-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[34] K. Caluwaerts *et al.*, "A biologically inspired meta-control navigation system for the psikharpax rat robot," *Bioinspiration Biomimetics*, vol. 7, no. 2, 2012, Art. no. 025009.

[35] M. Khamassi, P. Enel, P. F. Dominey, and E. Procyk, "Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters," *Progr. Brain Res.*, vol. 202, pp. 441–464, Jan. 2013.

[36] S. Palminteri, M. Khamassi, M. Joffily, and G. Coricelli, "Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters," *Nat. Commun.*, vol. 6, p. 8096, Aug. 2015.

[37] L. Kocsis and C. Szepesvári, "Discounted UCB," in *Proc. 2nd PASCAL Challenges Workshop*, 2006, pp. 784–791.

[38] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 235–256, 2002.

[39] G. Velentzas, C. Tzafestas, and M. Khamassi, "Bio-inspired meta-learning for active exploration during non-stationary multi-armed bandit tasks," in *Proc. IEEE Intell. Syst. Conf.*, London, U.K., 2017, pp. 661–669.

[40] L. Schilbach *et al.*, "Minds made for sharing: Initiating joint attention recruits reward-related neurocircuitry," *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2702–2715, 2010.

[41] R. Ooko, R. Ishii, and Y. I. Nakano, "Estimating a user's conversational engagement based on head pose information," in *Intelligent Virtual Agents* (Lecture Notes in Computer Science), vol. 6895, H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson, Eds. Heidelberg, Germany: Springer, 2011, pp. 262–268.

[42] R. Allesiardo and R. Féraud, "EXP3 with drift detection for the switching bandit problem," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Paris, France, Oct. 2015, pp. 1–7.

[43] A. Steinfeld *et al.*, "Common metrics for human–robot interaction," in *Proc. 1st ACM SIGCHI/SIGART Conf. Human–Robot Interact.*, 2006, pp. 33–40.

[44] E. Ferreira and F. Lefevre, "Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 256–274, 2015.

[45] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artif. Intell.*, vol. 166, nos. 1–2, pp. 140–164, 2005.

[46] G. Viejo, M. Khamassi, A. Brovelli, and B. Girard, "Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning," *Front. Behav. Neurosci.*, vol. 9, p. 225, Aug. 2015.

Mehdi Khamassi received the ECE degree from the Conservatoire National des Arts et Métiers, Paris, France, in 2003 and the D.E.A., Ph.D., and Habilitation degrees from Université Pierre et Marie Curie, Paris, in 2003, 2007, and 2014, respectively.

He is a Permanent Research Scientist with the French National Center for Scientific Research, Institute of Intelligent Systems and Robotics, Sorbonne Université, Paris. He is also a Visiting Researcher with the Intelligent Robotics and Automation Laboratory, National Technical University of Athens, Athens, Greece, and the Department of Experimental Psychology, University of Oxford, Oxford, U.K. He currently serves as the Director of Studies for the Cogmaster Program with the École Normale Supérieure/École des Hautes Études en Sciences Sociales/Université Paris Descartes, Paris. His current research interests include decision-making, reinforcement learning, performance monitoring and reward signals during social and nonsocial paradigms.

Dr. Khamassi currently serves as an Associate Editor for *Frontiers in Neurorobotics* and *Intellectica*.

George Velentzas is currently pursuing the ECE degree with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, with specialization in the fields of signals, control and robotics, computing systems, electronics and biomedical engineering.

He is currently a member of the research group with the Institute of Communications and Computer Systems researching on Core Robotic Functionality of BabyRobot EU-H2020 project by integrating adaptation and developmental learning schemes to optimize a behavior-based control architecture with cognitive feedback provided by human action in human–robot interaction scenarios, mainly at the lower level of a multiarmed bandit framework. His current research interests include robotics and machine learning with emphasis in reinforcement learning and fast adaptation in dynamic environments, complementing the approach with a computational neuroscience perspective.

Theodore Tsitsimis received the ECE degree from the School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, in 2018, focusing on robotics, machine learning, and electronics with a thesis on human–robot interaction including handover and joint attention tasks.

His current research interests include artificial intelligence, robotics, and computer vision.

Costas Tzafestas (M'99) received the ECE degree from the National Technical University of Athens (NTUA), Athens, Greece, in 1993, and the D.E.A. and Ph.D. degrees in robotics from the Université Pierre et Marie Curie (Paris 6 University), Paris, France, in 1994 and 1998, respectively.

In 2003, he joined NTUA School of ECE, where he currently serves as an Associate Professor in advanced robotics. He has been the Scientific Manager of the MOBOT Project (FP7) and is currently the Technical Manager of the I-SUPPORT Project and the Coordinator of the BabyRobot Project (H2020). His current research interests include cognitive assistive robotics, human/robot interaction, telerobotics and haptics, also spanning robust, adaptive and intelligent robot control and robot learning methods with applications in advanced robotic manipulation, as well as in walking and mobile robots. He has co-authored over 100 scientific publications and has participated in several national and international research programs.

Dr. Tzafestas currently serves as a Senior Editor for the *Journal of Intelligent and Robotic Systems*.