An extended framework for robot learning during child-robot interaction with human engagement as reward signal

M. Khamassi[1,2] and G. Chalvatzaki[1] and T. Tsitsimis[1] and G. Velentzas[1] and C. Tzafestas[1,3]

***Abstract*— Using robots as therapeutic or educational tools for children with autism requires robots to be able to adapt their behavior specifically for each child with whom they interact. In particular, some children may like to be looked into the eyes by the robot while some may not. Some may like a robot with an extroverted behavior while others may prefer a more introverted behavior. Here we present an algorithm to adapt the robot's expressivity parameters of action (mutual gaze duration, hand movement expressivity) in an online manner during the interaction. The reward signal used for learning is based on an estimation of the child's mutual engagement with the robot, measured through non-verbal cues such as the child's gaze and distance from the robot. We first present a pilot joint attention task where children with autism interact with a robot whose level of expressivity is pre-determined to progressively increase, and show results suggesting the need for online adaptation of expressivity. We then present the proposed learning algorithm and some promising simulations in the same task. Altogether, these results suggest a way to enable robot learning based on non-verbal cues and to cope with the high degree of non-stationarities that can occur during interaction with children.**

Keywords: HRI, Reinforcement Learning, Active Exploration, Autonomous Robotics, Engagement, Joint Action.

I. Introduction

This paper is an extension for the BAILAR 2018 workshop of the short paper accepted as Late Breaking Report at the RO-MAN 2018 main conference. We present recent progresses in developing robot learning abilities for the adaptation to human-specific requirements during child-robot interaction. In particular, we aim at enabling the robot to vary the level of expressivity of its actions in order to increase the child's mutual engagement with the robot and thus contribute to further develop children's social interaction skills.

Mutual engagement can be defined as "the process by which interactors start, maintain and end their perceived connection to each other during an interaction" [1]. More specifically, according to [2], "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". As is clear from these definitions, engagement is a wide notion which encompasses many different aspects and features. Nevertheless, here for sake of simplicity we will focus on measurements of joint attention between interacting child and robot which can be measured through gaze and body posture.

Researches in the field of social robotics have recently shown a growing interest in monitoring human and robot gaze during social interaction [3], [4], [5], [6]. 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 [7], [8], [9], [10], [11], [12], [13]. However, to our knowledge no one has yet proposed a way to make the robot learn on the fly which actions to perform in response to changes in human engagement. Previous researches having applied reinforcement learning to human-robot interaction have most of the time employed discrete action spaces (*e.g.* [14], [15], [16]), hence preventing generalization to more complex tasks requiring continuous motor actions.

In this work, we develop a robot reinforcement learning algorithm which uses human engagement monitoring signals as a reward signal during non-verbal social interaction. Specifically, 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). A second originality consists in applying the *parameterized* framework of reinforcement learning [17], [18] to human-robot interaction (HRI): this employs a set of discrete actions $A_d = \{a_1, a_2, ..., a_k\}$, where each action $a \in A_d$ features m_a continuous parameters $\{\theta_1^a, ..., \theta_{m_a}^a\} \in \mathbb{R}^{m_a}$, 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. The third originality of the proposed approach consists in achieving robot fast adaptation during social interaction through active exploration [19], [20], [21], [22]. The proposed solution relies on a novel combination of existing methods applied to a simple human-robot interaction scenario in the following manner: We apply Gaussian exploration [23] 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 [24], which tracks online variations of the agent's performance measured by short-term and long-term reward running averages. At each

[1]All authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece.

[2]Mehdi Khamassi is also with Sorbonne Université, CNRS, Institute of Intelligent Systems and Robotics, Paris, France.

[3]Correspondence: ktzaf@cs.ntua.gr

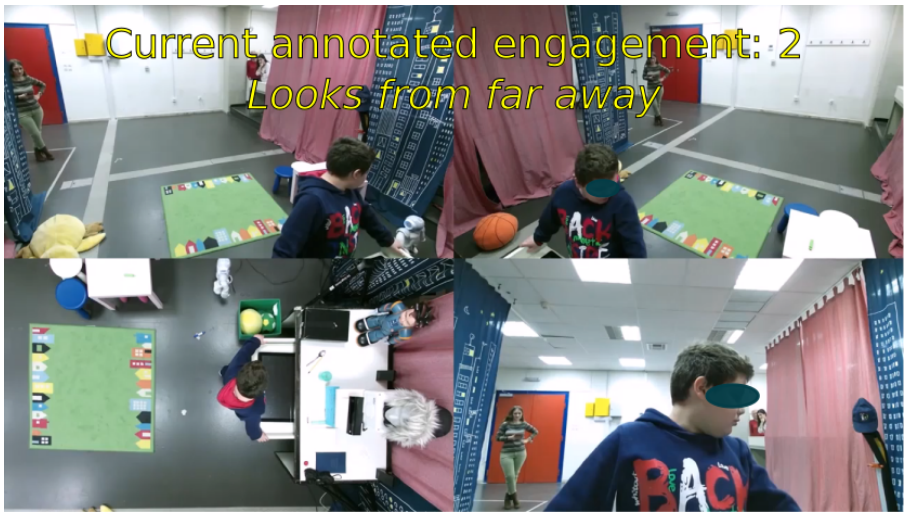


Fig. 1. Pilot child-robot interaction study with children with autism. The figure shows a moment where a child with ASD showed moderate engagement while the robot moved its arm up and down to point at an object on a table.

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 present paper starts by presenting the child-robot interaction pilot studies we have done with children with Autistic Spectrum Disorders (ASD), in which we illustrate the need for robot's adaptation to each specific child requirement. We then present the proposed reinforcement learning algorithm and its formalism. We finish by presenting a set of numerical simulations which are meant to assess its performance in simulation before deploying it during the real child-robot interaction experiments.

II. Pilot child-robot interaction study with children with autism

The general experimental paradigm adopted here consists in having a small humanoid robot interact with children (one at a time), under the supervision of an observing human adult, and finding the appropriate robot behavior to maximize children's engagement in the task. This paradigm follows the objectives defined in the framework of the EU-funded project BabyRobot (H2020-ICT-24-2015-6878310), where a set of child-robot interaction use-cases have been designed and implemented to study the development of specific socio-affective, communication and collaborative skills in typically developing children as well as children with ASD. In this framework, we have set up a pilot experiment[1] where the NAO robot is interacting with a child (Fig. 1), and repeatedly points at an unreachable object while varying the level of expressivity of its pointing gesture (i.e., opening-and-closing hand for a certain duration, bending its torso with a certain angle in the direction of the object, gazing at the child for a certain duration) until the child understands the "intention" of the robot and engages himself/herself into joint action in order to help the robot grasp the object. The engagement estimation, in this pilot study, was provided in real-time by an expert who observed the child during the interaction with the robot, considering five discrete levels of engagement (0 to 4, with 0 meaning absence of engagement and 4 meaning full engagement and attempt to offer help).

We present here some preliminary results for this real HRI task for which we have yet performed the experiment only with a small number of children with mild and moderate ASD symptoms, plus a few children with severe symptoms (12 children in total so far). First, children with severe symptoms expressed no interest in the task, neither in the condition with the robot nor in a control condition where the child interacts with a human expert rather than with the robot. In contrast, children with mild symptoms displayed great enthusiasm and interest in playing with the robot as well as with the researcher and enjoyed the whole process. These children were able to respond quite well to the task and completed the experiment with success. Overall, we found that two out of eight children with mild symptoms successfully maximized their engagement in joint attention with the robot and gave the object to the robot spontaneously. The remaining six children successfully increased their engagement, although not optimally, ending up moving the object closer to the robot but not handing it in. The two children with moderate symptoms also increased engagement and ended up exploring the object pointed at by the robot. Finally, again children with severe symptoms did not respond to the task.

Figure 1 shows one child performing the task, looking at the NAO robot (moderate engagement) while the latter moved its arm down after pointing at the object on the small white table. The psychologist, who can be seen near the red door, is manually annotating the child's engagement so that the robot can adapt its behavior. These results are promising

[1]This experiment has been approved by the ethical committee of Athena Research Center, Greece. The children's parents provided written consents.

and stimulating in that eight children that we interviewed after the task said that they would like to play more often with the robot and that they found the tasks we proposed them relatively easy. But many more subjects for each level of severity of ASD symptoms are required before allowing some statistics on the results. Interestingly, studying how the robot's movements affected the child's engagement, we observed that when the robot opened and closed its grip or exchanged glances between the child and the object for a period of time while pointing at the object, it contributed to an increase in the child's engagement. This suggests that varying the level of expressivity in the robot's actions in time was key to increase child engagement. Nevertheless, different levels of expressivity appeared to be appropriate for different children. It is thus relevant to propose a way for the robot to autonomously learn the appropriate degree of expressivity appropriate for each child.

Algorithm 1 Active exploration with meta-learning

Initialize $V_0(s)$, $\theta^a_{i,0}(s)$, $Q_0(s,a)$, β_0 and σ_0
for $t = 0, 1, 2, ...$ **do**
  Select discrete action a_t (Eq. 2)
  Select action parameters $\widetilde{\theta}^a_{i,t}$ (Eq. 3)
  Observe new state and reward (Eq. 6)
  Update $Q_{t+1}(s_t, a_t)$ (Eq. 1)
  Update $V_{t+1}(s_t)$ and $\theta^a_{i,t+1}(s_t)$ (Eq. 4-5)
  if meta-learning **then**
    Update reward running averages $\bar{r}_t$ and $\bar{\bar{r}}_t$
    Update β_{t+1} and σ_{t+1}
  end if
end for

III. Robot learning algorithm

The proposed algorithm is summarised in Algorithm 1. It is based on reinforcement learning with *parameterized* action spaces [17], [18]. It employs a set of discrete actions $A_d = \{a_1, a_2, ..., a_k\}$, where each action $a \in A_d$ features m_a continuous parameters $\{\theta^a_1, ..., \theta^a_{m_a}\} \in \mathbb{R}^{m_a}$, 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. Learning the value of discrete action $a_t \in A_d$ selected at timestep t in state s_t is done through Q-Learning [25]:

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

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:

$$P(a|s_t, \beta_t) = \frac{exp\left(\beta_t Q_t(s_t, a)\right)}{\sum_{a'} exp\left(\beta_t Q_t(s_t, a')\right)} \tag{2}$$

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

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

$$P(\widetilde{\theta}^a_{i,t}|s_t, a_t, \sigma_t) = \frac{1}{\sqrt{2\pi}\sigma_t} exp\left(-(\widetilde{\theta}^a_{i,t} - \theta^a_{i,t}(s_t))^2/(2\sigma_t^2)\right) \tag{3}$$

where the width σ_t of the Gaussian is tuned through meta-learning (see below) and continuous action parameters $\theta^a_{i,t}(s_t)$ are learned with the CACLA algorithm [23]. 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 critic and the actor:

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_C \delta_t \tag{4}$$

$$\theta^a_{i,t+1}(s_t) = \theta^a_{i,t}(s_t) + \alpha_A \delta_t (\widetilde{\theta}^a_{i,t} - \theta^a_{i,t}(s_t)) \tag{5}$$

where α_C and α_A are learning rates.

In order to perform active exploration, we apply a noiseless version of the meta-learning algorithm of [24], which tracks online variations of the agent's performance measured by short-term $\bar{r}_t$ and long-term $\bar{\bar{r}}_t$ reward running averages (with timeconstant τ_1 and τ_2 respectively). 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, and the width σ_t of the Gaussian distribution from which each continuous action parameter θ^a_i is sampled around its current value. The main idea is that when the performance is better than average, exploration should be decreased in order to reach optimality levels. In contrast, sudden drops in the performance should lead to increases in exploration in order to adapt to environmental non-stationarities.

Finally, we need to define a reward function for human-robot interaction tasks. This is not an easy task since during interaction the actions performed by a robot may have delayed effects on the human's behavior and on his engagement. To mimic this, we chose a reward component to be given by a dynamical system which is based on the virtual engagement E of the human in the task. This engagement 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 non-social extrinsic rewards such as food or money [26]. In our simulations, the quantified engagement arbitrarily starts at 5, increases up to a maximum $E_M = 10$ when the robot performs the appropriate actions with the appropriate parameters, and decreases down to a minimum $E_m = 0$ otherwise:

$$E_{t+1} = \begin{cases} E_t + \eta_1(E_M - E_t)H(\theta^a_t), & \text{if } a_t = a^\star \ \& \ H(\boldsymbol{\theta}^a_t) \geq 0 \\ E_t - \eta_2(E_m - E_t)H(\theta^a_t), & \text{if } a_t = a^\star \ \& \ H(\boldsymbol{\theta}^a_t) < 0 \\ E_t + \eta_2(E_m - E_t), & \text{otherwise} \end{cases} \tag{6}$$

where $\eta_1 = 0.1$ is the increasing rate, $\eta_2 = 0.05$ is the decreasing rate, and $H(\boldsymbol{x})$ is the re-engagement function given by

$$H(\boldsymbol{x}) = 2\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^\star)^T(\boldsymbol{\Sigma}^\star)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}^\star)\right) - 1 \quad (7)$$

where $a^\star$, is the optimal discrete action, $\boldsymbol{\mu}^\star$ is the optimal parameter vector $\boldsymbol{\theta}^a_\star$ for the optimal action and $\boldsymbol{\Sigma}^\star$ is a diagonal matrix $\sigma^{\star 2}\boldsymbol{I}$ of size $m_{a^\star} \times m_{a^\star}$. To picture the idea, the parameters for which $H(\boldsymbol{x}) = 0$ define the boundaries of an $m^\star_a$-dimensional ball in the parameter space, inside which the engagement is increased. In general, each parameter might have difference tolerance, however for all the experiments we will be using a common $\sigma^\star = 10$, while all parameter values will be in [-100,100]. Fig. 2 depicts H-function in the case where the optimal action has only one continuous parameter.

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 n-tuple $(a_t, \theta^a_{1,t}, \theta^a_{2,t}, .., \theta^a_{m_a,t})$ performed by the robot.

IV. Simulations

Here we present two sets of numerical simulations: (1) an abstract task where the robot needs at each trial to choose between 6 different cubes on a table to point at ($a_1, a_2, .., a_6$ are the discrete actions), and at the same time to choose a continuous parameter θ between -100 and 100 which abstractly represents the expressivity of the pointing gesture. We consider that the optimal action tuple $(a*, \theta*)$ that the robot has to learn is stable during 1000 timesteps, making the simulated child engagement vary according to Equation 6. Then the optimal action tuple abruptly changes in an unsignalled manner – representing a change in the expectation by the simulated child, to which the robot should adapt. The task was chosen to fine tune the parameters of the algorithm which optimize performance and as a first assessment of the algorithm's adaptivity; (2) the second set of numerical simulations employ a task identical to the child-robot interaction pilot task described above. This second task was simulated in the virtual robot experimentation platform (V-REP). In the considered scenario, the NAO robot points at an object on a table with different degrees of action expressivity so as to catch the child's attention and thus increase mutual engagement.

We used the first task as a reference experiment to perform an exhaustive search of the parameters that permit the algorithm to reach its highest performance in cases of multiple successive abrupt changes in conditions (Fig. 3). The parameter-set which produced the best performance yielded an average engagement of 9.2 (the arbitrarily defined maximum being 10). Interestingly, the performance was not very sensitive of the values of parameters τ_1 and τ_2, while in the original article they were always chosen so that $\tau_1 = \tau_2$ [24]. Figure 4 shows the average and standard deviation of the simulated engagement obtained for 10 simulations of the task with the proposed active exploration algorithm compared to a passive exploration version of it (*i.e.,* where instruction 10 in Algorithm 1 is removed so that $\beta_t = 35$ and $\sigma_t = 19$ are fixed after having been obtained through parameter optimization). The blue curve shows the performance of the algorithm without active exploration, which adapts to each new condition but never exceeds a plateau of about 6. This is because the algorithm continues to explore regularly even after having found the optimal action and continuous parameter to perform, as any algorithm with fixed exploration does. Obviously, an algorithm with an annealing process (iteratively decreasing exploration rate during a fixed duration) would have progressively reached optimal performance before the first change in task condition (timestep 1000). Nevertheless, it would not have been able to adapt to task changes. The red curve shows the performance of the algorithm with active exploration, which adapts faster and faster after each task change, only performing short transient

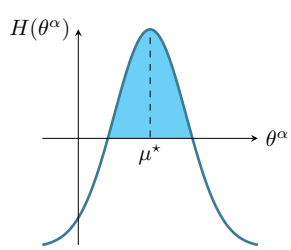


Fig. 2. Principle adopted to simulate variations of child engagement as a function of the distance between the robot's current continuous parameters of action and optimal ones.

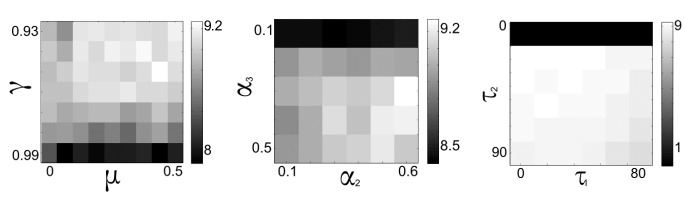


Fig. 3. Parameter optimization in the first task. Each datapoint corresponds to the average engagement obtained for 10 simulations of the task with a given parameter set. The color indicates engagement value between 0 (min) and 10 (max).

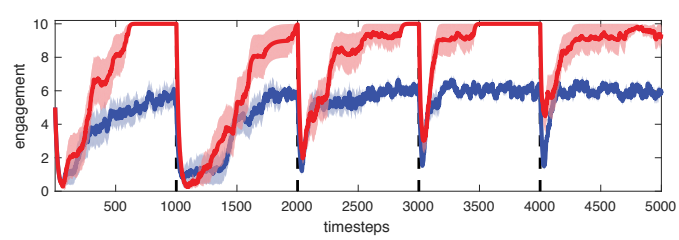


Fig. 4. Comparison of engagement in the first task for 10 simulations of the algorithm with active exploration (red), and a version of the algorithm without active exploration (blue). Vertical dashed lines represent changes in the optimal action tuple $(a*, \theta*)$ that the robot has to relearn every 1000 timesteps.

explorations when drops in performance are detected, and reaches each time the optimum engagement of 10.

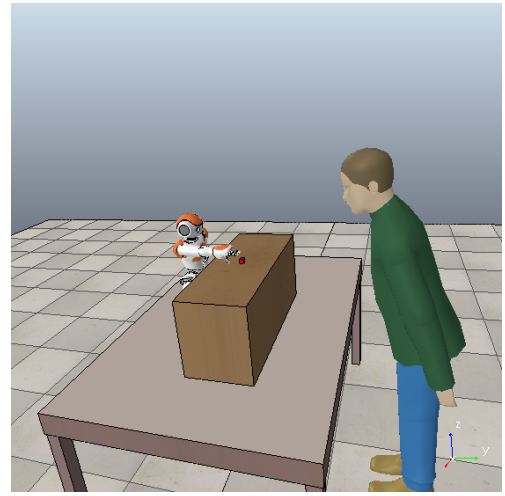
Fig. 5(b) illustrates how the algorithm works in the second task, which mimics the real child-robot pilot study described above. We parameterized the simulated pointing action of the robot with two parameters (t_1, t_2) corresponding to the time in seconds the robot would spend iteratively opening-closing its hand during pointing, and the time spent exchanging glances with the child. Examples of different expressivity levels defined by these parameters are shown in Table I.

	Pointing gesture
expressivity ↑	point + open-close + glance ($t_1 \neq 0, t_2 \neq 0$)
	point + exchange glance ($t_1 = 0, t_2 \neq 0$)
	point + open-close hand ($t_1 \neq 0, t_2 = 0$)
	point ($t_1 = 0, t_2 = 0$)

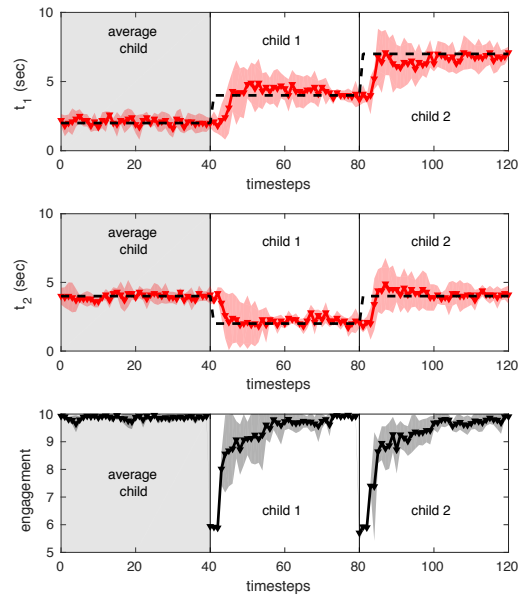
TABLE I

ROBOT'S POINTING ACTION IN THE SECOND TASK, WITH PARAMETERS CORRESPONDING TO INCREASING LEVELS OF EXPRESSIVITY.

We initialized the algorithm based on the parameters obtained on average during previous interactions with simulated children. This way, the algorithm started from a meaningful average value of action parameters/durations $(\bar{t_1}, \bar{t_2})$, rather than being initialized randomly, and then adapted to each specific child. We defined a time range from 0 to 10 seconds. Fig. 5(b) shows the average performance over 10 simulations. The robot firstly interacted with an "average child", meaning that the child engaged optimally with parameters $(\bar{t_1}, \bar{t_2})$. Then, at timestep 40, the experiment involved another child (child 1) with different optimal parameters. The engagement of child 1 was initially low but progressively re-increased as the robot was finding the optimal continuous action parameters. The figure also illustrated the increased variance in executed action parameters during exploration followed by a re-focus around the learned parameters during exploitation. Similarly, at timestep 80 child 2 took the place of child 1 and the robot readjusted its parameters. Importantly, we observe that in less than 10 timesteps the robot found the optimal parameter values for the different children whose engagement reached 8 in just a few timesteps. This thus illustrates a sufficiently fast adaptation process to work online during real child-robot interactions.



(a) V/REP SIMULATION ENVIRONMENT



(b) SIMULATION RESULTS

Fig. 5. Numerical simulations in the second task. (a) Setup used for the simulations of the same task as the pilot real child-robot interaction experiment. (b) Simulation results. **Left:** Before timestep 40 the robot executed the default parameters values, no adaptation was performed. After timestep 40, the robot adapted its action parameters (black) towards the optimal action parameters (red). **Right:** Child's engagement reached 90% within less than 10 trials.

V. CONCLUSIONS AND FUTURE WORK

In this short paper, we presented recent progresses in developing robot learning abilities for the adaptation to human-specific requirements during child-robot interaction. In particular, we aimed at enabling the robot to vary the level of expressivity of its actions in order to increase the child's mutual engagement with the robot and thus contribute to further develop children's social interaction skills. We first showed some preliminary results in a pilot study involving a robot with a predetermined sequence of increased expressivity of action while pointing at an unreachable object until a child with ASD understands that the robot needs help and engages in joint action. The preliminary results suggest that the level of expressivity does play a role in engaging the child, but should nevertheless be adapted through online learning to each interacting child. We then presented a learning algorithm based on reinforcement learning in parameterized action spaces [17], [18] – 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 – to which we added active exploration so as to cope with the frequent non-stationarities that can occur during human-robot interaction. We presented simulation results showing that the algorithm can adapt in a sufficiently small number of trials to be applied

to adaptation in real-time during interaction.

In future work, we plan to test the learning algorithm during real child-robot interaction. We moreover plan to study whether the average parameters over different interacting children is efficient or whether there exists distinct clusters of parameters – especially within the data obtained in the real experiments – that should be used as separate initialization points for the learning algorithm.

Acknowledgements

The authors would like to thank psychologists Christina Papaeliou and Asimenia Papoulidi, and the Special Elementary School for Children with ASD of Piraeus, Greece, for enabling us to make these experiments with children with ASD. This research work has been partially supported by the EU-funded Project BabyRobot (H2020-ICT-24-2015, grant agreement no. 687831), and by the Centre National de la Recherche Scientifique (MI ROBAUTISTE & PICS 279521).

References

[1] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005.

[2] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the Association for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[3] J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey, and J. Ventre-Dominey, "I reach faster when i see you look: gaze effects in human–human and human–robot face-to-face cooperation," *Frontiers in neurorobotics*, vol. 6, 2012.

[4] S. Al Moubayed, G. Skantze, and J. Beskow, "The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction," *International Journal of Humanoid Robotics*, vol. 10, no. 01, p. 1350005, 2013.

[5] H. Admoni, A. Dragan, S. S. Srinivasa, and B. Scassellati, "Deliberate delays during robot-to-human handovers improve compliance with gaze communication," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 49–56.

[6] S. Bampatzia, V. Vouloutsi, K. Grechuta, S. Lallée, and P. F. Verschure, "Effects of gaze synchronization in human-robot interaction," in *Conference on Biomimetic and Biohybrid Systems*. Springer, 2014, pp. 370–373.

[7] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 2010, pp. 375–382.

[8] D. Szafir and B. Mutlu, "Pay attention!: designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2012, pp. 11–20.

[9] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, 2015.

[10] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "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," *International Journal of Social Robotics*, vol. 9, no. 1, pp. 63–86, 2017.

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

[12] T. Ahmed and A. Srivastava, "A prototype model to predict human interest: Data based design to combine humans and machines," *IEEE Transactions on Emerging Topics in Computing*, 2017.

[13] A. Di Nuovo, D. Conti, G. Trubia, S. Buono, and S. Di Nuovo, "Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability," *Robotics*, vol. 7, no. 2, p. 25, 2018.

[14] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.

[15] M. Khamassi, S. Lallée, P. Enel, E. Procyk, and P. Dominey, "Robot cognitive control with a neurophysiologically inspired reinforcement learning model," *Frontiers in Neurorobotics*, vol. 5:1, 2011.

[16] E. Senft, P. Baxter, J. Kennedy, S. Lemaignan, and T. Belpaeme, "Supervised autonomy for online learning in human-robot interaction," *Pattern Recognition Letters*, vol. 99, pp. 77–86, 2017.

[17] W. Masson and G. Konidaris, "Reinforcement learning with parameterized actions," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[18] M. Hausknecht and P. Stone, "Deep reinforcement learning in parameterized action space," in *International Conference on Learning Representations (ICLR 2016)*, 2016.

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

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

[21] C. Moulin-Frier and P. Oudeyer, "Exploration strategies in developmental robotics: a unified probabilistic framework," in *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, 2013, pp. 1–6.

[22] F. Benureau and P.-Y. Oudeyer, "Behavioral diversity generation in autonomous exploration through reuse of past experience," *Frontiers in Robotics and AI*, vol. 8, 2016.

[23] H. van Hasselt and M. Wiering, "Reinforcement learning in continuous action spaces," in *IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 2007, pp. 272–279.

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

[25] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[26] L. Schilbach, M. Wilms, S. Eickhoff, S. Romanzetti, R. Tepest, G. Bente, N. Shah, G. Fink, and K. Vogeley, "Minds made for sharing: Initiating joint attention recruits reward-related neurocircuitry," *Journal of Cognitive Neuroscience*, vol. 22, no. 12, pp. 2702–2715, 2010.