Human-Guided Learning of Social Action Selection for Robot-Assisted Therapy

Emmanuel Senft, Paul Baxter, Tony Belpaeme
Centre for Robotics and Neural Systems,
Cognition Institute Plymouth University, U.K.
{emmanuel.senft, paul.baxter, tony.belpaeme}@plymouth.ac.uk

Abstract

This paper presents a method for progressively increasing autonomous action selection capabilities in sensitive environments, where random exploration-based learning is not desirable, using guidance provided by a human supervisor. We describe the global framework and a simulation case study based on a scenario in Robot Assisted Therapy for children with Autism Spectrum Disorder. This simulation illustrates the functional features of our proposed approach, and demonstrates how a system following these principles adapts to different interaction contexts while maintaining an appropriate behaviour for the system at all times.

1 Introduction

Humans are interacting increasingly with machines, and robots will be progressively more important partners in the coming years. Human-human interactions involve high dimensionality signals and require complex processing: this results in a large quantity of data that ideally needs to be processed by an autonomous robot. One potential solution is the application of machine learning techniques. Specifically, online learning is desirable, however, some level of initial knowledge and competencies are required to avoid pitfalls in the early phases of the learning process, particularly in contexts where random exploration could lead to undesirable consequences.

In this paper, we propose an approach inspired by, but separate from, learning from demonstration to guide this early learning of an action selection mechanism for autonomous robot interaction with a human, by taking advantage of the expert knowledge of a third-party human supervisor to prevent the robot from exploring in an inappropriate manner. We first present the formal framework in which our action selection strategy learning takes place (section 2), then illustrate this with a case study from the domain of Robot Assisted Therapy for children with Autism Spectrum Disorder (ASD), where the incorrect selection of actions can lead to an unacceptable impact on the goals of the interaction (section 3).

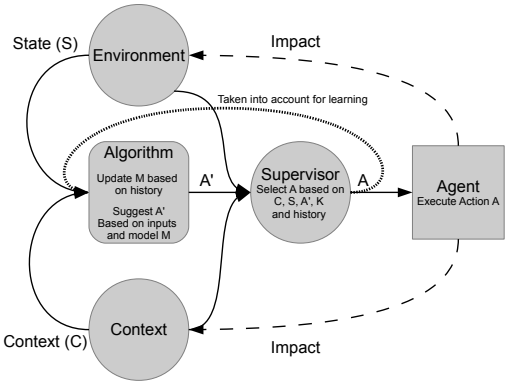

Figure 1: The supervised online learning of autonomous action selection mechanism.

2 Supervised Emergent Autonomous Decision Making

2.1 Framework

The situation considered involves a robotic agent, a human supervisor of the agent, and a human with which the agent, but not the supervisor, should interact. The agent proposes actions that are accepted or rejected by the supervisor prior to executing them. The method proposed in this paper aims at enabling

Appearing in Proceedings of the 4^{th} Workshop on Machine Learning for Interactive Systems (MLIS) 2015, Lille, France. JMLR: W&CP volume 43.

the agent to progressively and autonomously approximate the ideal behaviour as specified by the supervisor.

Our framework has five components: an agent and an environment interacting with each other, a supervisor, the algorithm controlling the agent and a context characterising the interaction between the agent and the environment. The agent has a defined set of available actions $\mathcal{A}$. The environment could be a human, a robot, or a computer for example and is characterised by a n-dimensional vector $\mathcal{S} \in \mathcal{R}^n$, which is time varying. The context $\mathcal{C} \in \mathcal{R}^m$ gives a set of parameters defining higher-level aspects, such as goals or the state, of the interaction, see figure 1. The supervisor and the agent have direct access to the context, but may ignore the real value of the state and see it only through observations.

The principal constraints are that the interaction has one or more high level goals, and some available actions can have a negative impact on these goals if executed in specific states. This should be avoided, so algorithms depending on randomness to explore the environment state space are inappropriate.

In order to simplify the system, we are making the following assumptions. Firstly, that the environment, while dynamic, is consistent: it follows a defined set of rules $\mathcal{E}$ which also describe how the context is updated. Secondly, the supervisor $\mathcal{T}$ is omniscient (complete knowledge of the environment), constant (does not adapt during the interaction), and coherent (will react with the same action if two sets of inputs are identical). Finally, the supervisor has some prior knowledge of the environment $\mathcal{K}$.

The algorithm has a model $\mathcal{M}$ of the supervisor and the environment and will update it through online learning following the learning method $\mathcal{L}$. $\mathcal{M}$ is iteratively updated based on supervisor feedback to approximate $\mathcal{T}$ and $\mathcal{E}$, in this way progressively approximating the action that the supervisor would have chosen, and what impact this would have on the environment. Equation 1 describes the update of each part of the framework from the step n to $n+1$.

$$\begin{aligned} M_n &: (C_{0\to n}, S_{0\to n}, A_{0\to n-1}, A'_{0\to n-1}) \longrightarrow A'_n \\ \mathcal{T} &: (C_{0\to n}, S_{0\to n}, A'_{0\to n}, A_{\to n-1}, \mathcal{K}) \longrightarrow A_n \\ \mathcal{E} &: (C_{0\to n}, S_{0\to n}, A_{0\to n}) \longrightarrow (S_{n+1}, C_{n+1}) \\ \mathcal{L} &: (C_{0\to n+1}, S_{0\to n+1}, A'_{0\to n}, A_{0\to n}) \longrightarrow M_{n+1} \end{aligned} \tag{1}$$

At the start of the interaction, the environment is in a state S_0 with the context C_0 and the algorithm has a model M_0. Applying M_0 to C_0 and S_0, the algorithm will select an action A'_0 and propose it to the supervisor. The supervisor can either accept this action or select a new one according to $\mathcal{T}$, and makes the agent execute the resulting action A_0. The environment will change to a new state S_1 and the context will be updated to C_1 according to $\mathcal{E}$. Based on S_1, S_0, C_1, C_0, A_0, and A'_0, the algorithm will update its model to M_1. The process can then be repeated based on the updated model.

2.2 Related Work

The approach we take here necessarily requires the application of machine learning, but we do not commit at this stage to a single algorithmic approach; the specific requirements for our application include online learning, deferring to an external supervisor, and being able to handle a dynamic environment.

A widely used method to transfer knowledge from a human to a robot is Learning from Demonstration (LfD), see [2] for a survey. In the case of policy learning, a teacher provides the learning algorithm with correct actions for the current state and repeats this state-action mapping for enough different states to give the algorithm a general policy. LfD is usually combined with supervised learning: trying directly to map outputs and inputs from a teacher, see [12] for a list of algorithms that can be used in supervised learning. The other important point is how the demonstrations are generated, a first approach is using batch learning: the teacher trains the algorithm during a training phase after which the robot is used in full autonomy [11]. Or there may be no explicit training phase; using online learning the demonstrations are given during the execution if required: the robot can request a demonstration for the uncertain states, e.g. when a confidence value about the action to perform is too low [6].

Another method is Reinforcement Learning: the algorithm tries to find a policy maximising the expected reward [3, 10]. However, this implies the presence of a reward function, which may not be trivial to describe in domains (such as social interaction) that do not lend themselves to characterisation. Consequently Abbeel and Ng proposed to use Inverse Reinforcement Learning by using an expert to generate the reward function [1], subsequently extended to use partially-observable MDPs [8], although expert-generated rewards also pose problems on the human side [17].

The goal of our proposed approach differs from these alternative existing methods. The intention is to provide a system that can take advantage of expert human knowledge to progressively improve its competencies without requiring manual intervention on every interaction cycle. This is achieved by only asking the human supervisor to intervene with corrective infor-

mation when the proposed action of the robot agent is deemed inappropriate (e.g. dangerous) prior to actual execution. This allows the robot to learn from constrained exploration; a consequence of this is that the load on the supervisor should reduce over time as the robot learns. The supervisor nevertheless retains control of the robot, and as such we characterise the robot as having *supervised autonomy*. Contrary to the active learning approach used by, for example, Cakmak and Thomaz [5] the robot is not asking a question and requiring a supervisor response, it is proposing an action which may or may not be corrected by the supervisor.

3 Case Study: Application to Therapy

One potential application area is Robot Assisted Therapy (RAT) for children with Autism Spectrum Disorders (ASD). Children with ASD generally lack typical social skills, and RAT can help them to acquire these competencies, with a certain degree of success, e.g. [7, 14]. However, these experiments typically use the Wizard of Oz (WoZ) paradigm [13], which necessitates a heavy load on highly trained human operators.

We propose the use of supervised autonomy [15, 16], where the robot is primarily autonomous, but the therapeutic goals are set by a therapist who maintains oversight of the interaction. Having a supervised autonomous robot would reduce the workload on the therapist.Both the therapist and the robot would be present in the interaction, the robot interacting with the child and the therapist supervising the interaction and guiding the robot while it is learning its action selection policy.

The formalism described above (section 2.1) can be directly applied to this scenario. In this case, the context is the state of the task selected by the therapist to help the child develop certain social competencies, for example, a collaborative categorisation game [4] intended to allow the child to practice turn taking or emotion recognition. The state may be defined using multiple variables such as motivation, engagement, and performance exhibited by the child during the interaction, the time elapsed since the last child's action, and their last move (correct or incorrect). The robot could have a set of actions related to the game, such as proposing that the child categorises an image, or giving encouragement to the child.

In this scenario, the goal would be to allow the child to improve their performance on the categorisation task, and this would be done by selecting the appropriate difficulty level and finding a way to motivate the child to play the proposed game. We can expect the child to react to the robot action and that these reactions can be captured by the different variables that define the child's state (as provided by therapists for example). In principle, while precise determinations are likely to be problematic, we assume that some aspects of these variables can be estimated using a set of sensors (e.g. cameras and RGBD sensors to capture the child's gaze and position; the way the child interacts with the touch screen; etc). For the remainder of this paper, however, we assume that a direct estimation of internal child states are available to the system.

3.1 Proof of concept

A minimal simulation was constructed to illustrate the case study described above. The state $\mathcal{S}$ is defined using three variables: the child's performance, engagement and motivation in the interaction. The robot has the following set of actions $\mathcal{A}$: encouragement (give a motivating feedback to the child), waving (perform a gesture to catch the child's attention), and proposition (inviting the child to make a classification). In this minimal example, the environment $\mathcal{E}$ is the child model. A minimal model of the child was constructed that encompassed both processes that were dependent on the robot behaviour (e.g. responding to a request for action), and processes that were independent of the robot behaviour (e.g. a monotonic decrease of engagement and motivation over time independently of other events). The reaction of the model follows a rule-based system, but the amplitude of the response is randomly drawn from a normal distribution to represent the stochastic aspect of the child's reaction and the potential influence of non-defined variables in the state. A number of simplifications are necessary, such as the assumption of strict turn-taking and interactions in discrete time. The minimal child model is summarised in figure 2.

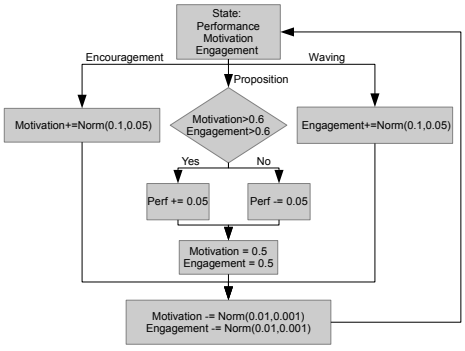

Figure 2: Model of the child used in the minimal simulation; random numbers are drawn from normal distributions.

Formally, the minimal simulation follows the framework established above (equation 1), with the simplification that a history of prior states, contexts, and

actions is not used in the learning algorithm. This results in a setup where the system makes a suggestion of an action to take, which the supervisor can either accept or reject, in which case an alternative action is chosen (figure 3).

This allows the supervisor to take a more passive approach when the algorithm selects an acceptable action since they will only have to manually select a corrective action when this is needed. If the learning method is effective, the number of corrective actions should decrease over time, decreasing the workload on the therapist over the interaction.

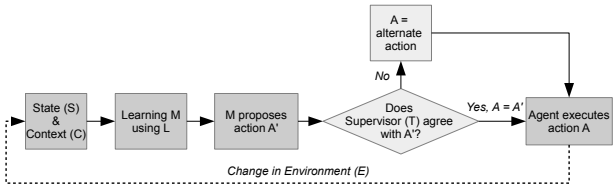

Figure 3: Description of agent's action selection process: the agent proposes actions that are validated by the supervisor prior to execution.

The learning model $\mathcal{M}$ is a MultiLayer Perceptron (MLP), with three input nodes for the input states, three output nodes for the three possible actions and nine nodes in the hidden layer. The model is trained using backpropagation (as $\mathcal{L}$), the true labels are given by the supervisor decision: output of 1 for the action selected by the supervisor (A) and -1 for the other ones. A Winner-Takes-All process is applied on the output of the MLP to select the action suggested by the robot (A').

Figure 4 shows a subset of a run from step 100 to 150. With this approach, there is no distinct learning and testing phases, but in the first part of the interaction (before step 100), the supervisor had to produce multiple corrective actions to train the network to express the desired output. The strategy used by the supervisor is the following: if the motivation is lower than 0.6 the supervisor enforces the action 'encouragement', else if the engagement is below 0.6 'waving' is enforced, and if both are above 0.6 then a proposition is made. The first graph presents the evolution of the state over time, and the second one the output of the MLP for each action. The vertical red lines represent an intervention from the supervisor, i.e. a case where the supervisor enforces a different action than the one suggested by the MLP. The action actually executed is represented by a cross with the same colour as the respective curves.

Figure 5 shows a comparison of the cumulative total of the different actions suggested and of the intervention as well as the child performance for three different models of a child and for a random action selection scheme. The difference in the child models in the three first graphs is the value of the thresholds required for a good classification action, high reactive child: 0.6 and 0.6, asymmetrically: 0.9 for encouragement and 0.6 for engagement, and low reactive: 0.9 and 0.9. Below these thresholds, a proposition would lead to a bad action decreasing the performance. It can be observed that the algorithm learns different strategies for each child and that there is more learning apparent at the start of the interaction than at the end (the rate of interventions is decreasing over time), indicating that the system is choosing the appropriate action at the appropriate time, and that the workload on the supervisor (necessity to provide these corrective actions) decrease over time. The last plot demonstrates a random action selection with a high reactive child. Contrary to the other cases, the child's performance decreases over time, and the number of interventions increases. Here, a *bad* action only decreases the performance, but in reality it may result in the termination of the interaction, which must be avoided.

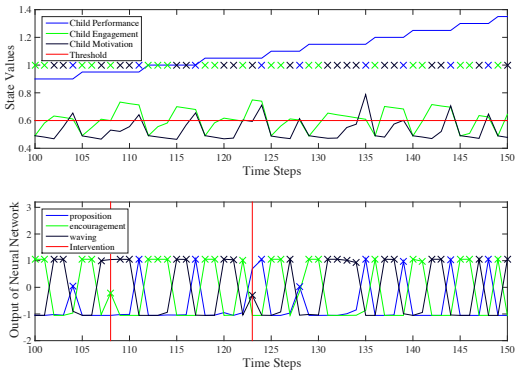

Figure 4: Subset of an interaction from step 100 to 150.

4 Discussion

While demonstrating promise, there are a number of limitations to the framework as presented. The assumptions described in section 2.1 are typically violated when working with humans. Firstly the children are all different, and a method learned for one child may often not be suited when working with another. Furthermore, the same child may have non-consistent behaviour between the sessions and even within a single session. There is no perfect solution to solve this problem, but we can expect that with enough training sessions and a more complex learning algorithm, the system would be able to capture patterns and react to the different behaviours appropriately. Since it is ex-

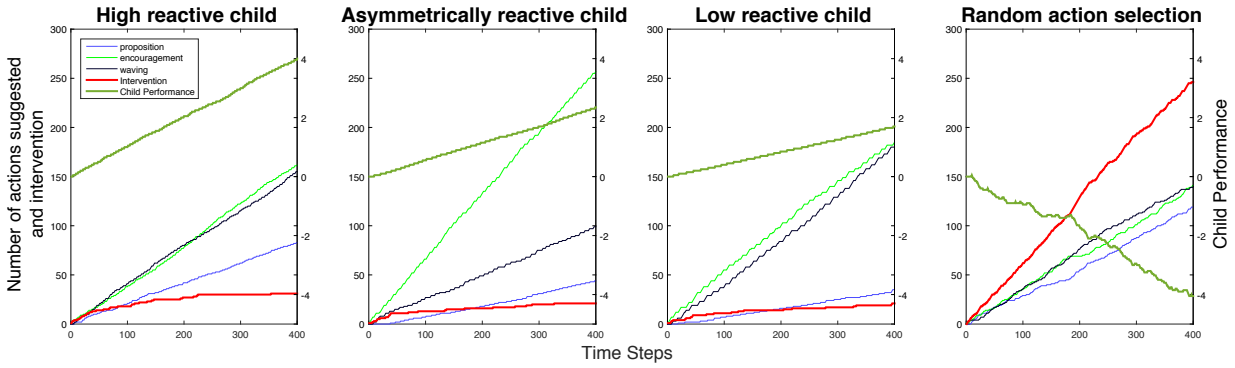

Figure 5: Comparison of the cumulative total of the different actions suggested, the supervisor interventions required, and child performance for three different models of a child (highly responsive; asymmetrically increased responsiveness to engagement than motivation; low responsiveness), and a random action selection scheme.

pected that in a real application of such an approach a therapist who knows the child will always be present, we propose that for a new child the algorithm will use a generic strategy based on previous interactions with other children, with subsequent fine-tuning under supervision.

Another assumption that is likely to be violated is that of a perfect supervisor. As explained in [6] humans are not always consistent nor omniscient, but methods presented in the literature can be used to cope with these inconsistencies if enough data is gathered. Further mitigating solutions could be employed, such as the robot warning the therapist if it is about to select an action which had negative consequences in a previous interaction (even if for a different child). Furthermore, it may not be possible to measure the true internal states of the child in the real world, with imperfect estimations of these states being more likely accessible. In this case, inspiration from [9] can be used to mixed the POMDP framework with the help of an exterior oracle. Another problem which will have to be addressed in the future is the difference in inputs between the robot and the therapist: the therapist will have access to language, more subtle visual features and their prior experience, whereas the robot may have direct and precise access to some aspects of the child's overt behaviour (such as timings of touchscreen interaction).

In the currently implemented case study, we assume that the supervisor responds to the action proposed by the robot within some predetermined fixed time, whether this response is accept or reject (figure 3). This, in principle, allows the supervisor to only actively respond if a proposed action is clearly inappropriate. In further developments, we will incorporate a measure of certainty (given prior experience) into the time allowed the supervisor to respond to the proposed action: for example increasing the time available if the confidence in the proposed action is low. This modulation of the load on the supervisor's attention according to confidence should result in the supervisor being able to increasingly pay attention to the child directly, rather than to the robot system, as training progresses.

5 Conclusion

We have presented a general framework to progressively increase the competence of an autonomous action selection mechanism that takes advantage of the expert knowledge of a human supervisor to prevent inappropriate behaviour during training. This method is particularly applicable to application contexts such as robot-assisted therapy, and our case study has provided preliminary support for the utility of the approach. While the simulation necessarily only provided a minimal setup, and thus omitted many of the complexities present in a real-world setup, we have nevertheless shown how the proposed method resulted in the learning of distinct action selection strategies given differing interaction contexts, although further refinement is required for real-world application. Indeed, given real-world supervisor knowledge limitations, we suggest it will furthermore be possible for a suitably trained action selection mechanism of this type to aid the supervisor in complex and highly dynamic scenarios.

Acknowledgements

This work is supported by the EU FP7 DREAM project (grant 611391).

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 1–8, 2004.

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[3] A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[4] P. Baxter, R. Wood, and T. Belpaeme. A touchscreen-based sandtrayto facilitate, mediate and contextualise human-robot social interaction. In *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pages 105–106. IEEE, 2012.

[5] M. Cakmak and A. L. Thomaz. Designing robot learners that ask good questions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 17–24. ACM, 2012.

[6] S. Chernova and M. Veloso. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research*, 34(1):1, 2009.

[7] K. Dautenhahn. Robots as social actors: Aurora and the case of autism. In *Proc. CT99, The Third International Cognitive Technology Conference, August, San Francisco*, volume 359, page 374, 1999.

[8] F. Doshi, J. Pineau, and N. Roy. Reinforcement learning with limited reinforcement: Using bayes risk for active learning in pomdps. In *Proceedings of the 25th international conference on Machine learning*, pages 256–263. ACM, 2008.

[9] F. Doshi, J. Pineau, and N. Roy. Reinforcement learning with limited reinforcement: Using bayes risk for active learning in pomdps. In *Proceedings of the 25th international conference on Machine learning*, pages 256–263. ACM, 2008.

[10] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, pages 237–285, 1996.

[11] W. B. Knox, S. Spaulding, and C. Breazeal. Learning social interaction from the wizard: A proposal. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[12] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[13] L. Riek. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 1(1):119–136, Aug. 2012.

[14] B. Robins, K. Dautenhahn, R. T. Boekhorst, and A. Billard. Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105–120, July 2005.

[15] E. Senft, P. Baxter, J. Kennedy, and T. Belpaeme. When is it better to give up?: Towards autonomous action selection for robot assisted asd therapy. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, pages 197–198, New York, NY, USA, 2015. ACM.

[16] S. Thill, C. A. Pop, T. Belpaeme, T. Ziemke, and B. Vanderborght. Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn*, 3(4):209–217, Apr. 2013.

[17] A. L. Thomaz and C. Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, 2008.