SPARC: an efficient way to combine reinforcement learning and supervised autonomy.

Emmanuel Senft
Centre for Robotics and Neural Systems
Plymouth University
Plymouth, PL4 8AA, United Kingdom
emmanuel.senft@plymouth.ac.uk

Séverin Lemaignan
Centre for Robotics and Neural Systems
Plymouth University
Plymouth, PL4 8AA, United Kingdom
severin.lemaignan@plymouth.ac.uk

Paul E. Baxter
Lincoln Centre for Autonomous Systems
University of Lincoln
Lincoln, LN6 7TS, United Kingdom
pbaxter@lincoln.ac.uk

Tony Belpaeme
Centre for Robotics and Neural Systems
Plymouth University
Plymouth, PL4 8AA, United Kingdom
iMinds – Ghent University
Department of Electronics and Information Systems
B-9052 Ghent, Belgium
tony.belpaeme@plymouth.ac.uk

Abstract

Shortcomings of reinforcement learning for robot control include the sparsity of the environmental reward function, the high number of trials required before reaching an efficient action policy and the reliance on exploration to gather information about the environment, potentially resulting in undesired actions. These limits can be overcome by adding a human in the loop to provide additional information during the learning phase. In this paper, we propose a novel way to combine human inputs and reinforcement by following the Supervised Progressively Autonomous Robot Competencies (SPARC) approach. We compare this method to the principles of *Interactive Reinforcement Learning* as proposed by Thomaz and Breazeal. Results from a study involving 40 participants show that using SPARC increases the performance of the learning, reduces the time and number of inputs required for teaching and faces fewer errors during the learning process. These results support the use of SPARC as an efficient method to teach a robot to interact with humans.

1 Introduction

To be widely used by non-technical people, robots have to be able to learn, in order to adapt their behaviour to new challenges and tasks. These robots have to acquire knowledge whilst interacting in an environment which possibly includes other people. Reinforcement Learning [14] is a machine learning algorithm specifically designed to address the issue of learning how to interact efficiently based on feedback from the environment. This learning method has already been widely applied to robots [10], however, as pointed by Knox and Stone in [9], the reward function from the environment can either not be defined for certain tasks or at least be sparse in its assignation of reward. A solution is to include a human in the learning process, moving from classical machine learning to interactive machine learning. In this framework, a human supervisor is fully integrated in the learning process and can provide additional information to the algorithm to improve the learning [4]. Furthermore, this approach also provides end users with the ability to steer the learning in the direction they desire, which can improve the robot's usability [1].

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Multiple approaches have been proposed to combine human feedback and reinforcement learning. In [8], Knox et al. present the TAMER framework, designed to teach an action policy in the absence of any environmental feedback using a human to provide the missing rewards used for the reinforcement learning. Thomaz and Breazeal [16] propose to combine human and environmental rewards and use them directly as input for a Q-Learner. During their experiments, they observed that participants tried to use rewards as a way to guide the robot's actions. Consequently, they introduced a second guidance channel to guide the robot action in follow-up studies and observed better learning.

However, we argue that the lack of control over the robot's action in these methods limits the impact of the human in the learning loop. By taking inspiration from Learning from Demonstration [2, 3], the human can provide demonstration of the desired action policy, and at the same time interactively teach the robot. Following this approach, we have proposed the Supervised Progressively Autonomous Robot Competencies – SPARC [13]. This is based on the supervised autonomy framework [15]: the robot can act autonomously, but a human is supervising it to prevent undesirable actions from being executed if necessary. By adding machine learning (reinforcement learning in this case), in which the robot can learn from the human corrections and improve its action policy over time whilst only executing actions deemed appropriate by the supervisor.

This paper presents the combination of SPARC and reinforcement learning and compares it with a previously applied approach following the principles of *Interactive Reinforcement Learning* [16] using four metrics: performance, teaching time, number of inputs and risks taken while teaching. We show that in each of these metrics, SPARC leads to a significant improvement over *Interactive Reinforcement Learning*, supporting its use as a framework to teach robots in an interactive fashion in sensitive environments, such as those typically encountered in human robot interaction.

2 Methodology

Problem specifications In this paper, we tackle the action selection problem in an environment modelled as a deterministic Markov Decision Process. An agent can execute actions changing the current state to a new one according to a fixed deterministic transition function. A limited number of states provide rewards (positive or negative), and the agent has to maximise the rewards obtained over time. Additionally, a human supervisor is present and can provide additional information to the robot to improve the learning (rewards and guidance for IRL or commands for SPARC).

Interactive Reinforcement Learning Due to its clarity, simplicity and aim to be used for human-robot interaction, the method used as a benchmark in this paper is *Interactive Reinforcement Learning* (IRL) following principles proposed in [16]. Thomaz and Breazeal proposed a first example of incorporating a human in the learning process by directly combining the reward from the environment with human rewards: a human supervisor can provide rewards which are combined to the environmental ones and used with Q-Learning. Following early studies, authors enriched the interaction with three mechanisms to improve the teaching: a guidance mechanism to direct the robot's attention to some actions (without covering the entire action space), a communication of uncertainty and an undo behaviour executing an action cancelling the previous one after a negative reward. This study uses an algorithm inspired from the one proposed by Thomaz and Breazeal and implementing these additions.

Supervised Progressively Autonomous Robot Competencies In [12] and [13], we proposed the Supervised Progressively Autonomous Robot Competencies (SPARC) as an interaction framework allowing a supervisor (human or other) to teach a robot an action policy. SPARC is centred around a suggestion/correction mechanism whereby the agent suggests actions to its supervisor which can either correct the action by selecting another one or let the action be executed after a short delay by not reacting to the suggestion. This system allows the supervisor to be totally in control of the actions being executed by the robot. A learning algorithm, here reinforcement learning, learns from the supervisor decision to improve the suggestions over time, decreasing the necessity of correcting actions and thus reducing the workload on the supervisor.

As the supervisor only provides commands and no reward to the robot, this approach is initially not designed to be used with reinforcement learning. However the constant control of the supervisor on the robot's action implies that every action executed by the robot has been implicitly or actively validated by the supervisor, so all these executed actions can receive a positive reward: we reward 0.5 for actions actively selected by the supervisor, and 0.25 for the actions passively accepted.

Evaluation task To evaluate the efficiency of combining SPARC with reinforcement learning and to compare the results with IRL, we run a study using Sophie's kitchen, the initial setup used by Thomaz and Breazeal in [16]. Participants have to teach a virtual robot how to bake a cake in the environment presented in Figure 1. The robot can pick-up, drop or use objects and can move left or right between three locations: the shelf, the table and the oven. Six main steps have to be completed to bake the cake: placing the bowl on the table, putting a first ingredient in the bowl (flour or eggs), then the other one, mixing with the spoon, emptying the bowl in the tray and finally putting the tray in the oven. As shown later, we used these six steps to evaluate participants' performance. As argued by Thomaz and Breazeal, this environment is interesting for interactive learning due to the large number of states (more than 10,000), multiple non-trivial successful action policies (minimum of 28 actions to achieve the goal) and success and failure states used to provide environmental rewards: for example, if the spoon is put in the oven, a failure state is reached, providing a negative reward, ending the current teaching episode and returning the environment to the initial state. More detailed information can be found in [16]. This environment has been reimplemented to be used in this study, and the two interaction methods are using strictly the same learning algorithm, only the way to interact changes.

(a) Step 0: Initial state. (b) Step 3: ingredients in the bowl. (c) Step 6: Success.

Figure 1: Sophie's kitchen, the environment used in the study in three different states.

Study setup The study involves 40 participants (age M=25.6, SD=10.09; 24F/16M) divided into two groups. The first group interacts with IRL and the second one with SPARC. Participants first teach the robot how to complete the task and then a testing phase, where participants' inputs are disabled, evaluates the robot behaving on its own to assess participants performance in teaching. To limit the study time, a hard limit of 25 minutes for the teaching phase has been set for both systems, but participants could move on to the testing whenever they desired.

This paper presents a subset of the results of a larger study having each participant interacting three times with each system. The full results are currently being analysed.

3 Results

As not all participants reached the goal state (i.e., a cake in the oven) during training, the performance is expressed on a scale from 0 to 6 representing how many of the 6 main steps presented in section 2 (putting the bowl on table, adding an ingredient...) are autonomously completed by the robot in the testing phase. Results on four metrics (performance, interaction time, number of failure during teaching and number of inputs given by the teacher) are presented in Table 1 and Figure 2.

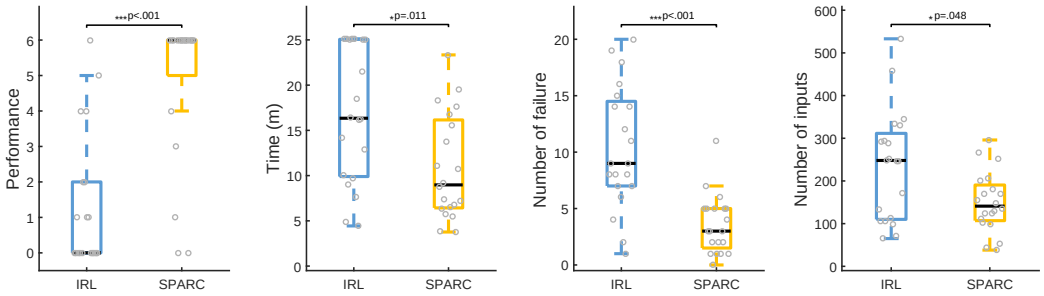

Figure 2: Comparison of the performance, interaction time, number of failures and number of inputs for the two conditions. Black horizontal bars represent medians and grey circles raw data points.

Table 1: Results of evaluation metrics for the two systems ($n = 20$). Using Wilcoxon rank sum test (results being non-normal), SPARC is significantly more efficient than IRL on all four metrics.

Metric	Median IRL	Median SPARC	$Z\text{-}value$	$p\text{-}value$	Effect size
Performance	0	6	-4.2	$< .001$	-0.67
Time (min)	16.3	8.97	2.53	0.011	0.4
Number of failure	9	3	4.06	$< .001$	0.641
Number of inputs	248	141	1.98	0.048	0.312

In this study, most of the participants using IRL did not reach a single step toward success (median of 0). This does not mean that this method cannot be used to teach an action policy: some participants reached the goal state with IRL and an expert would consistently achieve the goal state in this task. However, due to the more complex reward scheme and other challenges to interpret the trainers' rewards [5, 6, 11] not tackled by this method, participants need to have more in-depth understanding of how to interact with the algorithm to achieve success. These results support our thesis that relying only on feedback and guidance is a suboptimal method to teach a robot: even in this simple scenario, non-expert participants perform poorly. On the other hand, the median performance of 6 for SPARC shows that at least half of the participants reached the goal when interacting with SPARC and this in a shorter time, facing fewer failures during teaching and using fewer input. As such the combination of reinforcement learning and SPARC seems a more efficient teaching method.

4 Discussion

This study used a relatively simple environment: it has discrete states and a deterministic transition function. Realistic environments will be more complex and challenging. Furthermore, this simulation did not contain human interactants, but interacting with people adds two major constraints; unlike simulation we cannot train the agent for a long time before obtaining a correct action policy. In addition, as soon as the robot is used with people its behaviour has to be appropriate: suboptimal actions might have negative consequences. For this reason, the presence of a human supervisor having control over the robot's action has many advantages: it ensures that the behaviour expressed is appropriate and provides robustness against probabilistic environments, sensory errors and imperfect action policies. Whilst it is being controlled, the robot can progressively learn from the supervisor, and smoothly become more autonomous over time, reducing the workload on the supervisor. As the robot is acting in the real world and is executing a correct action policy, the need for exploration is reduced thus accelerating the learning process.

By relying on human commands and corrections, SPARC changes the teaching paradigm compared to classical interactive reinforcement learning methods. The human control over the robot's actions allows to bypass the need for users to manually assign rewards or evaluations to actions. It also uses only one-way feedbacks (selection) compared to two-way feedback in classical approaches (positive or negative reinforcement), thus preventing SPARC to face some of the challenges of human rewarding practices as described in [5, 11]. For example, a lack of feedback (absence of correction of an action) can either be direct passive support of the proposed action or if the action should have been corrected, it can be due to a slow reaction time or a desire not to interact. However, the control over the robot actions can allow supervisors to correct the trajectory when required and thus assuming a passive support in all cases where feedback is missing can still lead to efficient learning. SPARC could be combined with more classical Learning from Demonstration approaches [3] to teach lower level action policies, such as direct motor control where correct actions cannot be easily selected.

In this study, *Interactive Reinforcement Learning* achieved a poor performance with only a limited number of participants succeeding to use it to teach the robot how to bake the cake. On the other hand, SPARC achieved a high success rate in a shorter time and with fewer failures and lower teaching effort. This is consistent with [7], where authors argue that feedback channels are not an efficient method to teach an action policy from scratch, they recommend to start with Learning from Demonstration and then move to feedbacks for fine tuning. By relying on human intervention to prevent poor performance before it occurs, this paper has shown how SPARC can be usefully applied to teach an action policy while maintaining high performance, avoiding dangerous situations, and yet without overloading the human supervisor.

Acknowledgments

This work was supported by the EU FP7 DREAM project (grant no. 611391) and EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227).

References

[1] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[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. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. In *Springer handbook of robotics*, pages 1371–1394. Springer, 2008.

[4] J. A. Fails and D. R. Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45. ACM, 2003.

[5] S. Griffith, K. Subramanian, J. Scholz, C. Isbell, and A. L. Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2625–2633, 2013.

[6] C. L. Isbell Jr, M. Kearns, S. Singh, C. R. Shelton, P. Stone, and D. Kormann. Cobot in lambdamoo: An adaptive social statistics agent. *Autonomous Agents and Multi-Agent Systems*, 13(3):327–354, 2006.

[7] T. Kaochar, R. T. Peralta, C. T. Morrison, I. R. Fasel, T. J. Walsh, and P. R. Cohen. Towards understanding how humans teach robots. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 347–352. Springer, 2011.

[8] W. B. Knox and P. Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16. ACM, 2009.

[9] W. B. Knox and P. Stone. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 5–12. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[10] J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 2013.

[11] R. Loftin, B. Peng, J. MacGlashan, M. L. Littman, M. E. Taylor, J. Huang, and D. L. Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous Agents and Multi-Agent Systems*, 30(1):30–59, 2016.

[12] E. Senft, P. Baxter, and T. Belpaeme. Human-guided learning of social action selection for robot-assisted therapy. In *4th Workshop on Machine Learning for Interactive Systems*, 2015.

[13] E. Senft, P. Baxter, J. Kennedy, and T. Belpaeme. SPARC: Supervised progressively autonomous robot competencies. In *International Conference on Social Robotics*, pages 603–612. Springer, 2015.

[14] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[15] 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, Journal of Behavioral Robotics*, 3(4):209–217, 2012.

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