From Evaluating to Teaching: Rewards and Challenges of Human Control for Learning Robots

Emmanuel Senft[1], Séverin Lemaignan[2], Paul Baxter[3] and Tony Belpaeme[1,4]

***Abstract*— Keeping a human in a robot learning cycle can provide many advantages to improve the learning process. However, most of these improvements are only available when the human teacher is in complete control of the robot's behaviour, and not just providing feedback. This human control can make the learning process safer, allowing the robot to learn in high-stakes interaction scenarios especially social ones. Furthermore, it allows faster learning as the human guides the robot to the relevant parts of the state space and can provide additional information to the learner. This information can also enable the learning algorithms to learn for wider world representations, thus increasing the generalisability of a deployed system. Additionally, learning from end users improves the precision of the final policy as it can be specifically tailored to many situations. Finally, this progressive teaching might create trust between the learner and the teacher, easing the deployment of the autonomous robot. However, with such control comes a range of challenges. Firstly, the rich communication between the robot and the teacher needs to be handled by an interface, which may require complex features. Secondly, the teacher needs to be embedded within the robot action selection cycle, imposing time constraints, which increases the cognitive load on the teacher. Finally, given a cycle of interaction between the robot and the teacher, any mistakes made by the teacher can be propagated to the robot's policy. Nevertheless, we are are able to show that empowering the teacher with ways to control a robot's behaviour has the potential to drastically improve both the learning process (allowing robots to learn in a wider range of environments) and the experience of the teacher.**

I. INTRODUCTION

Interactive Machine Learning (IML) [1], [2] differs from Classical Machine Learning (CML) in the fact that the learning process is not one single monolithic step leading to a static classifier or robot behaviour, but a continuous iterative improvement of the behaviour. IML relies on a series of small learning steps progressively leading to a complete and autonomous system. Additionally, IML makes use of humans in the learning loop, to direct the learning process, making it at the same time faster, more adequate to the task and more efficient.

IML can take two forms: human supported classifiers (closer to semi-supervised learning) or agents learning to interact from human guidance (supervised reinforcement learning). A classical example of the first category is Active Learning, a learning process giving to the learner the opportunity to take a more active stance in the process, asking questions and querying labels from an oracle, often a human being [3]. The second category relates to agents learning to interact in an environment and profiting from humans inputs to improve the learning process. In this case, the learner is not in control of the datapoints it has to classify as those come directly form the environment; in fact, the agent interacts in an environment reacting to its actions and it requires a policy leading to a successful outcome in the task. The human can provide additional information to support the agent in developing its policy.

This work is focused on the second category, agents learning from human supervision to interact in an environment. An example is presented in Figure 1, where a robot is taught to interact with a child, supporting them in an educational activity. Compared to CML, IML holds the promises of faster and more flexible learning leading to a policy more adapted to current task [1], [2].

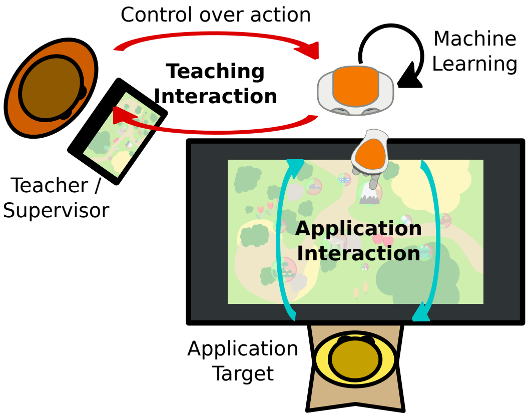

Fig. 1. Example of a human teaching a robot to interact with a child in an educational scenario.

In the context of agents learning to interact, a classical approach is to use a human to provide rewards on the robot's behaviour [4]. The scenario is similar to Reinforcement Learning (RL) [5], where an agent interacts in an environment providing rewards and where the agent has to maximise a notion a cumulative reward. Compared to traditional RL, using humans to distribute rewards possesses many advantages: no explicit reward function has to be provided, the human can anticipate the impacts of actions, reducing the challenge of credit assignment, and finally, the teacher can scaffold their reward distribution to help the agent to progressively improve its action policy [6]. This

*This work was supported by the EU FP7 DREAM project (grant no. 611391).

[1] University of Plymouth, CRNS, Plymouth, UK
emmanuel.senft@plymouth.ac.uk
[2] University of the West of England, BRL, Bristol, UK
[3] University of Lincoln, L-CAS, Lincoln, UK
[4] Ghent University, IDLabimec, Ghent, Belgium

way to support agent learning is attractive as it already provides advantages compared to classic RL and requires a simple interface between the teacher and a robot: the teacher only needs to be able to observe the robot's behaviour and provide a scalar evaluation of the learner's behaviour. However, as shown by [2], [7], [8], human teachers desire to have more control over the robot's behaviour and this control can improve drastically the learning.

This paper will present a definition of human control in the context of IML, as well as the advantages and challenges faced when applying it to teach robots or agents to interact in an environment. Throughout this paper, examples and results will be presented from a study exploring how a robot can be taught to support child learning in a educational task. The setup was presented in [9]. The study compared 3 conditions, a supervised robot interactively learning to support children, an autonomous robot re-enating the demonstrated policy and a passive robot providing no support to children and serving as a control condition. Final results are yet to be published.

II. Human Control

Robot learning possesses a unique opportunity compared to human learning in that the teacher can be fully in control of the learner's behaviour. This power over the learner provides many opportunities for agents learning from humans. Instead of simply providing feedback or labels as one would do for animals teaching for example [4], the teacher can actively decide the learner's behaviour, for example by demonstrating an efficient way of acting. Methods such as Learning from Demonstration (LfD) [10], [11] leverage this opportunity, often in manipulation scenario, to reach quickly an efficient behaviour. LfD has also been applied for interactive agents [12], [13], with offline learning. However, interactive learning with partial control for the teacher [7], [14], [15] hold significant promises as it would allow to deploy robots as blank slates and simply let the end user set the desired behaviour.

However, this partial control can be pushed further and we define 'human control' as the capacity for the teacher to ensure the robot executes a desired behaviour. This control can be achieved through a mixed-initiative control, where the robot behaves autonomously, while being supervised by the teacher and learning from this supervision. This semi-autonomous control needs to allow the teacher to select actions for the robot to execute, while letting the human prevent incorrect actions to negatively impact the world. This mixed-initiative control could for example follow the approaches proposed by [16] or [17], where a teacher can select actions for the robot to execute, and the robot can propose actions to the human. Depending on the method and the context, the proposition would be executed straightaway, with a short delay or only after approval by the teacher. Having the robot involved in the action loop might reduce the requirements on the teacher and the human in the loop ensures that the robot behaviour is correct at all time, even when the robot starts to learn to interact, a feature absent from methods such as RL.

In the study considered as example, the human control was provided using SPARC [17], a method allowing a teacher to select actions for a robot to execute. Based on these demonstrations, a learning algorithm creates a policy and each action is submitted to the teacher before an automatic execution. This allows the teacher to ensure that only useful actions are executed while not having to manually enforce each action required from the robot.

III. Advantages

This human control leads to several advantages compared to autonomous learning or feedback based teaching: the learning can be safe, fast and generalise more easily to different tasks. Additionally, trust can be built between the learner and teacher.

A. Safety

One of the main advantages of providing control over the robot's action to the teacher is safety. By ensuring that a human can prevent incorrect actions to have an impact on the real world, the policy executed by the agent is safer. This feature is especially interesting as many environments where artificial agents should be able to learn might present physical risks for the agent itself and surrounding humans, or risks of emotional harm. As the learner starts with an imperfect policy, incorrect actions are susceptible to be executed, but should be avoided at all costs. By providing control over the learner's actions to a human, such methods ensure a safe robot behaviour, thus increasing the range of environments where agents can learn and applications where they could be deployed.

In the study, the teacher could teach a robot *in-situ* an interactive policy to interact with children. Even in the first interactions, the teacher's oversight allowed the robot to display a behaviour suited to the interaction and supporting children in their learning task.

B. Speed

By indicating which actions an agent should take, a teacher can both lead the agent to an efficient policy and ensure the agent only explores parts of the environment that are relevant to the current task. Furthermore, if provided with an adequate interface, the teacher can provide the agent with additional details explaining the demonstrations or their choices, helping the learner to obtain more information about the environment than solely the demonstration. These three effects making a fuller use of the teacher, beyond simply labeling actions, can drastically quicken the learning process.

Despite learning only from 25 interactions with children (resulting in around 1 hour and half of teaching), in the study the teacher managed to inculcate the robot with a policy leading to a similar distribution of actions (cf Figure 2) and impact on the children in the autonomous and supervised conditions. It should be noted that as the interaction involved children, the resulting environment was non-repeatable, stochastic, social and sensitive; but despite these challenges, the results showed a successful teaching, demonstrating the efficiency of SPARC.

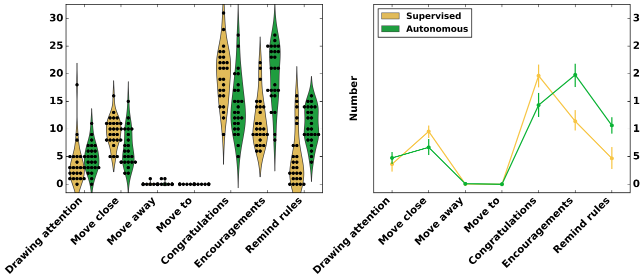

Fig. 2. Distribution of actions executed by the autonomous and the supervised robot.

C. Generalisation

Additionally, providing control to the teacher allows them to specify precisely the desired agent policy. This, combined with the faster learning would allow agents to learn policies tailored to a specific task from a generic definition of the world. This implies that robots could have access to a world representation with a large number of dimensions, allowing for a wide range of policies and tasks, and from this generic representation of the world, learn a policy directly suited to an application context.

Using guidance from the teacher, the algorithm created an efficient policy mapping a state in 210 dimensions to an action space composed of 655 discrete actions, thus demonstrating that from a large state and action spaces, this type of interaction allows to create a policy tailored to a specific task. Other tasks and policies could have been covered with the same representation of the world, interface and algorithm, but were not evaluated in that study.

D. Trust

By progressively teaching an agent to behave, a human teacher can build a model of the agent and create expectations on the agent's behaviour. This accumulated knowledge might lead to a trust between the teacher and the learner: by supervising the agent interacting in the world, the teacher can estimate the performance of the displayed policy. This trust and knowledge about the agent's capabilities might then ease its deployment to interact autonomously in the real world.

In a report written by the teacher while she was supervising the robot, she reported: "robot was often suggest[ing] good things" and "[I] Need to trust the robot more". In later post-study interviews she reported that she started to trust the robot in the last interactions, even if this trust never reached a level of full trust.

IV. Challenges

While giving human teachers control over the learner's actions provides advantages, it also raises challenges in the design of the interaction, the communication between the learner and its teacher, and in the application to specific time sensitive tasks.

A. Interface

The interface between the learner and the human teacher is key when designing and implementing IML applications. To provide enough control on the robot's behaviour ensure that the behaviour executed is safe for the agent and the surrounding partners, and reach an efficient policy, the teacher needs to be able to pre-empt any actions about to be executed by the robot before they negatively impact the environment. Additionally, the teacher needs to be able to select any action for the agent to execute. This implies that the interface needs at the same time to communicate the robot's intentions, allow the teacher to evaluate them and select actions to be executed if required.

Human-robot interactions rely on the robot displaying appropriate social behaviours, which requires often a large set of sensory inputs to interpret human behaviours and a large number of actions available to the robot. For example, in the study, the robot had access to 655 actions. Giving access to the teacher to such a large action space can be challenging. However, depending of the application, ways can be found to enable it. For example, for the study we used a Graphical User Interface (GUI) and we inferred the exact action selected by the teacher from her interaction with a representation of the world on the GUI instead of providing 655 buttons.

B. Human Time

Providing the robot's intentions to the teacher early enough to allow them to prevent actions to impact the world can be a challenge too, especially as some environments are time-critical. For example, a car driving semi-autonomously and detecting an obstacle requiring emergency breaking might not have the opportunity to wait for an explicit approval from the teacher. On the other an inappropriate emergency breaking is also highly dangerous as it would confuse and surprise other drivers. Consequently, the timing of actions and the way to ensure human oversight is a serious challenge when designing semi-autonomous agents.

A second challenge lies in the pace of the interaction. Today, a large part of the progress in ML relies on large quantities of data; however, when a human is included in the action loop, gathering data is a slow and tedious process. Even if datapoints arrive at 1Hz, the time required to accumulate the millions of examples required for methods such as Deep Learning [18] can be prohibitive (more than 250 hours). As such, systems relying on single humans to interactively provide demonstrations need data-efficient algorithms able to make better use of each datapoint.

The first challenge, time for reaction, can be mitigated by having different types of actions, corresponding to different ways of being communicated and approved. The second point was addressed in the study by requiring the teacher to specify features of the environment she used to select her actions. This additional information provided crucial details allowing the algorithm to make better use of demonstrations, learning a policy from only a limited number of demonstrations.

C. Human Limits

The last consideration is human limits. People are sensitive to workload and putting them under too much pressure will lead to human errors that will have to be corrected. When using human teachers, their workload needs to be minimised and ways need to be provided to recover from errors. This recovery needs to handle two sides: the learning algorithm needs to be informed about inaccurate demonstrations, and on the other hand, the impact of the erroneous actions on the environment needs to be corrected if possible. For example, a robot interacting with humans would need to be able to apologise in case of errors in order to maintain the trust surrounding humans have in it and allow the interaction to continue without friction.

In the study, the teacher reported herself making a few errors throughout the teaching process. She had access to a button to remove datapoints from the learning algorithm and thus correct the algorithm side of the error. However we didn't plan for error recovery in case of incorrect robot behaviour as we initially assumed the human behaviour would be constantly correct. In future implement, we will implement ways to recover from erroneous actions on the environment side too (such as apologies).

V. Discussion

The position defended in this paper is as follows:

> **When teaching robots to interact, human teachers should not be simply evaluating an autonomous behaviour, but should be able to control precisely the robot behaviour when required.**

The robotics and IML communities need to give a more complete role to the teachers, moving away from acting as simple oracles who label datapoints, and towards the incorporation of all facets of social learning, while taking advantage of the unique opportunities that artificial learners offer. More specifically, a learning robot should leverage people's natural skills at teaching humans and animals (transparency of the teaching process, scaffolding of the teacher's feedback/tasks and constant feedback from the learner), while also profiting from the features only available to artificial agents such as perfect memory, absence of tiredness or boredom, but especially the opportunity to control exactly the learner's behaviour.

Providing humans with this control can be a challenging task given the complexity of the problem. However, we contend that the gains outweigh these limitations dramatically compared to autonomous learning, learning from demonstration or retrospective evaluation of the robot's actions. Consequently, we suggest that research in HRI and IML should dedicate more effort towards this goal.

References

[1] J. A. Fails and D. R. Olsen Jr, "Interactive Machine Learning," in *Proceedings of the 8th International Conference on Intelligent User Interfaces*. ACM, 2003, pp. 39–45.

[2] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the People: The Role of Humans in Interactive Machine Learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[3] B. Settles, "Active learning literature survey," University of Wisconsin-Madison, Computer Sciences Technical Report 1648, 2009.

[4] 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*. ACM, 2009, pp. 9–16.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 1998.

[6] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive Learning From Policy-Dependent Human Feedback," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[7] A. L. Thomaz and C. Breazeal, "Teachable Robots: Understanding Human Teaching Behavior to Build More Effective Robot Learners," *Artificial Intelligence*, vol. 172, no. 6, pp. 716–737, 2008.

[8] 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.

[9] E. Senft, S. Lemaignan, M. Bartlett, P. Baxter, and T. Belpaeme, "Robots in the Classroom: Learning to Be a Good Tutor," in *4th Workshop on Robots for Learning (R4L) - Inclusive Learning, at HRI*, 2018.

[10] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A Survey of Robot Learning From Demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[11] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot Programming by Demonstration," in *Springer Handbook of Robotics*. Springer, 2008, pp. 1371–1394.

[12] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How to Train Your Robot-Teaching Service Robots to Reproduce Human Social Behavior," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium On*, 2014, pp. 961–968.

[13] P. Sequeira, P. Alves-Oliveira, T. Ribeiro, E. Di Tullio, S. Petisca, F. S. Melo, G. Castellano, and A. Paiva, "Discovering Social Interaction Strategies for Robots From Restricted-Perception Wizard-of-Oz Studies," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation*. IEEE Press, 2016, pp. 197–204.

[14] S. Chernova and M. Veloso, "Interactive Policy Learning Through Confidence-Based Autonomy," *Journal of Artificial Intelligence Research*, vol. 34, no. 1, 2009.

[15] W. Saunders, G. Sastry, A. Stuhlmueller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 2067–2069.

[16] T. Munzer, M. Toussaint, and M. Lopes, "Efficient Behavior Learning in Human-Robot Collaboration," *Autonomous Robots*, pp. 1–13, 2017.

[17] E. Senft, P. Baxter, J. Kennedy, and T. Belpaeme, "SPARC: Supervised Progressively Autonomous Robot Competencies," in *International Conference on Social Robotics*. Springer, 2015, pp. 603–612.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, 2015.