Mutual Reinforcement Learning with Robot Trainers

Sayanti Roy, Emily Kieson, Charles Abramson, Christopher Crick
Oklahoma State University
Stillwater, Oklahoma
sayanti.roy@okstate.edu, kieson@okstate.edu, charles.abramson@okstate.edu, chriscrick@cs.okstate.edu

***Abstract*—The researchers in this study have developed a novel approach using mutual reinforcement learning (MRL) where both the robot and human act as empathetic individuals who function as reinforcement learning agents for each other to achieve a particular task over continuous communication and feedback. This shared model not only has a collective impact but improves human cognition and helps in building a successful human-robot relationship. In our current work, we compared our learned reinforcement model with a baseline non-reinforcement and random approach in a robotics domain to identify the significance and impact of MRL. MRL contributed to improved skill transfer, and the robot was able successfully to predict which reinforcement behaviors would be most valuable to its human partners.**

I. Introduction

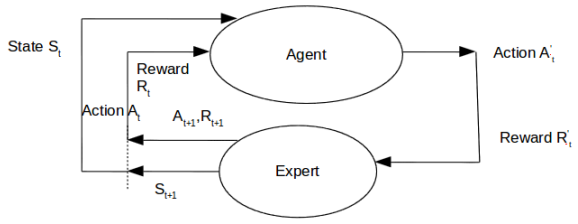

Figure 1. Mutual Reinforcement Learning

If instructional tasks could be deputized to robots, more students might be able to learn more skills more quickly. Thus to better facilitate natural social interactions and human learning, robots need to adapt and respond appropriately to each individual. In our MRL approach, robots are provided with a repertoire of specific approaches to human interaction based on principles of learning, and are able to adapt and respond in ways that are tailored based on an individual's unique responses. In this research, the humanoid robot Baxter is trying to identify human motivational preferences [1] (such as verbal encouragement, offering a reward, interacting with a positive gesture, or doing nothing at all) [2] and adapts accordingly to the situation, trading off exploration and exploitation [3] of these preferences in a skill transfer scenario while trying to establish a successful human-robot relationship.

II. Technical Description

In Algorithm 1, V is the set of weighted reinforcers to be given out during the task performance. For Baxter, $|V| = 4$,

This work was supported by NSF award #1527828 (NRI: Collaborative Goal and Policy Learning from Human Operators of Construction Co-Robots).

Algorithm 1 Mutual Reinforcement Learning

1: $V \leftarrow \{set\ of\ weighted\ items\}$
2: $R_i \leftarrow \{selected\ reinforcer\}$
3: **while** true **do**
4: $P_i(n_i) = w_i / w_{j_{\sum_{s_j \in V}}}$
5: $Weighted\ Randomly\ select\ V_i \in V\ according\ to\ P_i$
6: **if** A $\leftarrow$ success **then**
7: $V_i + \alpha$
8: $V_{s_j - V_i} - \ (s_j - 1)/\alpha$
9: **else**
10: $V_i - \alpha$
11: $V_{s_j - V_i} + \ (s_j - 1)/\alpha$
12: $\lambda_{s_j \in V} = (\lambda_{s_j} * \Phi) + (\lambda_{s_{j-1}} * (1 - \Phi))$
13: $R_i = \lambda_{s_j} + \sigma_{s_j}$
14: **end**

and R_i is the reinforcer selected by a weighted random choice where P_i is the probability or weight of each reinforcer R_i. An exponentially weighted moving average (EMWA) of each value is calculated over the previous five interaction iterations. We used this technique because the EMWA λ gives more importance to recent data. Hence the robot will make decisions based preferentially on the most recent human performance. Φ is the multiplier which keeps track of the latest five human robot interactions. A two-standard-deviation threshold is calculated and added to the EMWA to get the immediate variation *(line 13)* in the performance of the reinforcers. This will tell us that how the participant has responded when the distributions of the reinforcers are considered over α. After this, the weights of the reinforcers are recalculated for a new interaction. Thus after a motivation is given, if the candidate performes correctly, the weight of the reinforcer gets manipulated by α *(empirically set to 0.02 for this particular pattern making task)*. Motivations are only given out in the first phase of the experiment. To determine skill transfer success, the participants are later allowed to interact with the robot without any motivational assistance.

According to Fig. 1, MRL is a tuple $\{S, A(A'), T, R(R')\}$ where S is a set of states; A and A' are sets of actions; T is the set of state transition probabilities $P(s, a)$ upon taking action a in state s, and R and $R' : S \Rightarrow R(R')$ are the reward functions. Since, in MRL, the action of the agent is the reward to the expert and vice versa, the tuple can be simplified as follows: Agent=$\{S, A', T, R\}$, Expert=$\{S, A, T, R'\}$ where if

the agent executes action A', reward R' is received by the expert. This helps the expert to execute action A using an exploration/exploitation strategy, which at the same time acts as a reward R to the participant. If the action A' is successful, then the robot realizes that the participant is fonder of reward R, which acts at the same time as reward R' for the robot to understand its own performance or action A.

III. Experimental Procedure and Results

$n = 34$ (age μ=19.69, σ=3.49, male=13, female=21) participants were recruited for the experiment. Participants were asked to reconstruct a pattern from blocks initially taught by Baxter. At first, the pattern making involves positive reinforcers (again selected through MRL) from the robot if the participant fails at any point. They are asked to reconstruct the pattern twice later without reinforcers to observe the mental model in the skill transfer mechanism. The tasks were designed to cover perception, guided response, mechanism, adaptation etc. from Simpson's psychometric model [4] to analyze skill transfer from expert to novice. At the end of each experiment, participants' opinions were probed with questionnaires using a 5-point Likert scale.

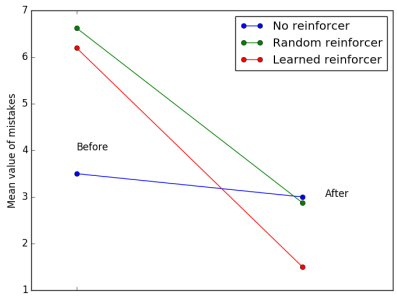

Figure 2. Task performance improvement. The "Learned reinforcer" in red used mutual reinforcement learning.

According to Simpson's psychometric model, the task which the motivations were attempting to assist is considered guided response. To determine the efficacy of the skill transfer scenario, the participants were asked, after a learning phase which incorporated motivational feedback, to repeat the same task without such motivations. In Fig. 2, the points denote the mean value of mistakes before and after the skill transfer method. We can clearly see that after the skill transfer phenomenon, the decline in the number of mistakes of the learned model is more than that of the other two groups.

With the random reinforcement group, the motivations provided by the robot were evenly distributed among reward (μ=2.5, σ=2.87), verbal (μ=3.22, σ=3.87), gesture (μ=3.4, σ=3.40), and none (μ=3, σ=3.23). In the case of learned models, the distribution was more skewed, as befits the attempt to match motivations to the preferences of each individual subject. Thus the distribution was reward (μ=2, σ=1.89), verbal (μ=1.3, σ=1.92), gesture (μ=1.91, σ=2.27), and none (μ=1.25, σ=1.28).

Table I
Participant response

Group	Play again		Good Teacher	
	M	**SD**	***M***	**SD**
None	3.54	0.69	4	0.63
Random	4	0.77	3.64	0.92
Learned	3.91	0.79	3.25	0.86

In Table I, the 5 point Likert scale corresponds to 1:Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly disagree. Overall 67% of the subject population thought that Baxter is a good teacher.

At the end of the experiment, those participants in the learned model group were asked to choose their preferred reinforcer. Within this group, out of 12 participants, two participants did not commit any mistakes and thus never received any reinforcement. Out of the remaining 10, however, Baxter could correctly identify the preferred reinforcers in five cases (twice as effectively as a random baseline). Thus, mutual reinforcement learning allowed the robot successfully to identify the cognitive orientation of the participants to a large extent.

IV. Conclusion and Future Work

At the state of the art, different social and cognitive development strategies are being developed which ensure not only better understanding of human cognition but also help humans and robots achieve a common goal through a collective set of actions while encouraging them throughout the process. In our future experiments we would like to incorporate an emotion recognition engine to identify subjects' emotions during the task, incorporating emotional responses into the robot's mutual reinforcement learning. In this way, we can determine the mental models arising out of this bidirectional learning policy, providing valuable information about how robots can train themselves over time to accomplish tasks and make necessary decisions with their human partners. Such models will give us a clearer idea about the heuristics responsible for the decisionmaking processes of both robots and humans. Among these other valuable real-world impacts, this research will enhance a robot's potential to help train and transfer skills in an apprenticeship learning context.

References

[1] J. Fasola and M. J. Mataric, "Using socially assistive human–robot interaction to motivate physical exercise for older adults," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2512–2526, 2012.

[2] S. Roy, E. Kieson, C. Abramson, and C. Crick, "Using human reinforcement learning models to improve robots as teachers," in *Proceedings of the 13th ACM/IEEE Conference on Human-Robot Interaction (HRI)*, 2018.

[3] B. Clement, D. Roy, P.-Y. Oudeyer, and M. Lopes, "Multi-armed bandits for intelligent tutoring systems," *Journal of Educational Data Mining*, vol. 7, no. 2, pp. 20–48, 2015.

[4] E. Simpson, "The psychomotor domain," *Washington DC: Gryphon House*, 1972.