MUTUAL REINFORCEMENT LEARNING TO IMPROVE ROBOTS AS TRAINERS

By

SAYANTI ROY

B.Tech in Electronics and Communication Engineering
West Bengal University of Technology
Kolkata,West Bengal
2014

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2020

MUTUAL REINFORCEMENT LEARNING TO

IMPROVE ROBOTS AS TRAINERS

Dissertation Approved:

Dr. Christopher Crick

Dissertation Adviser

Dr. Joe Cecil

Dr. Nohpill Park

Dr. Weihua Sheng

ACKNOWLEDGEMENTS

Its no secret that COVID-19 has a HUGE impact on the world and graduating during this time of immense crisis is a mixed feeling. In the past five years while working for my thesis I often use to imagine how the day of my graduation would look like, but destiny had a different plan it seems. In this tough time I would like to express my deep appreciation and thanks to a group of people who immensely supported and guided me throughout my journey in graduate school, my life and specially during the current situation. By great good fortune I cannot express how rewarding it is to work under the supervision of Dr. Christopher John Crick. Since my first day in graduate school, he believed in me like nobody else, gave me tremendous support and always encouraged my research ideas. It all started in Fall 2015 when he offered me a great opportunity to join the Robotics Cognition Laboratory. I also got the golden opportunity to be his teaching assistant for the undergraduate course he teaches and also served as his research assistant for five years. On the academic level, he taught me fundamentals of conducting scientific research in human robot interaction and machine learning. Under his supervision, I learned how to ask important research questions, find a solution to it, and finally publish the results. On a personal level, he inspired me by his disciplined, hardworking and passionate attitude. To summarize, I would be forever indebted to my advisor Dr. Crick for seeing something in me which I could not see in myself and making me the researcher I am today. Besides my advisor, I would like to thank the rest of my dissertation committee members (Dr. Joe Cecil, Dr. Nohpill Park and Dr. Weihua Sheng) for their great support and invaluable advice and my labmates for making my experience in Robotic cognition Laboratory fun and exciting. I would also like to thank National Science foundation (NSF) for the award 1527828 (NRI: Collaborative Goal and Policy Learning from Human Operators of Construction CoRobots) which helped in shaping my research. This dissertation would not have been possible without the intellectual contribution of my colleague Dr. Emily Kieson. I am thankful to her for her collaboration and contribution in conducting experiments and various research papers that helped me produce this dissertation. Last but not the least this journey would have been impossible without the endless support and love of my parents, my best friend Jitesh Krishnan and my other family members.

Name: SAYANTI ROY

Date of Degree: MAY, 2020

Title of Study: MUTUAL REINFORCEMENT LEARNING TO IMPROVE ROBOTS AS TRAINERS

Major Field: COMPUTER SCIENCE

Abstract: Recently, collaborative robots have begun to train humans to achieve complex tasks, and the mutual information exchange between them can lead to successful robot-human collaborations. In this thesis we demonstrate the application and effectiveness of a new approach called mutual reinforcement learning (MRL), where both humans and autonomous agents act as reinforcement learners in a skill transfer scenario over continuous communication and feedback. An autonomous agent initially acts as an instructor who can teach a novice human participant complex skills using the MRL strategy. While teaching skills in a physical (block-building) or simulated (Tetris) environment , the expert tries to identify appropriate reward channels preferred by each individual and adapts itself accordingly using an exploration-exploitation strategy. These reward channel preferences can identify important behaviors of the human participants, because they may well exercise the same behaviors in similar situations later. In this way, skill transfer takes place between an expert system and a novice human operator. We divided the subject population into three groups and observed the skill transfer phenomenon, analyzing it with Simpson”s psychometric model. 5-point Likert scales were also used to identify the cognitive models of the human participants. We obtained a shared cognitive model which not only improves human cognition but enhances the robots cognitive strategy to understand the mental model of its human partners while building a successful robot-human collaborative framework.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Daily experiences influence our learning and change the way we think and act. Sometimes we are not even aware that we are learning from our surroundings, which is a very informal way of perceiving things. On the other hand, we can also learn in a formal way from a a structured classroom environment. Learning is not limited to acquiring knowledge or facts; we also learn skills and attitudes. This can happen in different ways. We learn new ideas and concepts from a lecture or a discussion, whereas skills must be acquired via continuous practice and receiving simultaneous feedback from an instructor. In a planned environment, learning is reinforced by teachers who expect students to memorize the content and later reward them for it. In contrast, researchers or scientists learn by investigating things themselves, over time. But in any form of learning, motivations and rewards play a very important role, as people derive satisfaction from the feeling of competence. In the case of learning a new skill, people can be strongly motivated by the incentives they are given, which might lead them to acquire new knowledge which they can use in future.

Teachers are entrusted with the job of determining what the student will learn. They are not only instrumental to the students' learning, but also make sure that they have learnt the subject properly. Teaching must be planned very carefully, taking the learning styles and the background of the students into account. Teachers also need to assess students often to determine

how well they are progressing and simultaneously attend to their weaknesses. Hence teaching and learning are constructed over a series of intrinsic and extrinsic social interactions which influence the cognitive models of both the teacher and the learner. The learning can be improved if the facilitator can teach each individual, possessing some understanding of the subtleties of the student's mind, behavior and learning style, and regulating the motivational strategies accordingly (Prozesky, 2000).

Phase-I

In this work, we not only explore the use of building a human-understandable representation of a complex task through active learning and a conversational interface to improve the LfD process, but we explore the problem of robotic teaching as well. A set of American Sign Language (ASL) motions is taught to the humanoid robot Baxter, which learns to communicate the sentence "Hello, please listen to me" using its left arm. During the learning process, Baxter can request information from the human expert using label and demonstration queries. While learning, the robot is able to segment and hierarchically construct the components of the demonstrated task using expert feedback. This enables the human expert operator to understand the learning activity of the robot and jointly guide the teaching process, depending on the input the robot receives from the end-user. While learning from demonstration, the robot improves its learned model by interacting with its human teacher and cooperatively building a structure of hierarchical semantic labels (Kaelbling & Lozano-Pérez, 2011; Wu, Lenz, & Saxena 2014 ; Roy, Maske, Chowdhary, & Crick, 2017), as illustrated by our experimental results. The goal of the work described here is to create robots with sufficient human-accessible task understanding so that they can act as successful tutors or coaches for complex skill learning. Compared to experts, novice learners are often unsuccessful at noticing important information and task patterns. They fail to prioritize subtasks in terms of their importance to the overall goal, and hence tend to attend inappropriately to distracting or irrelevant aspects of the task environment. If the entire task can be segmented into well-organized subtasks with meaningful labels

(Figure 1.), the novice should be better able to distinguish relevant relationships, and should learn better. In this work we quantified the performance of Baxter as a learner and a teacher using the root-mean square error (RMSE) of the robots arm trajectories, compared to a standard task demonstration. When Baxter is employed as a learner, the RMSE is calculated comparing the task success under traditional LFD versus query-driven active learning which jointly builds a semantic structure with a human expert. Participants are recruited to act as experts and novice human operators. The participants allotted in the novice group are taught by Baxter with and without this semantic scaffolding. The RMSE is calculated comparing the performances of these two groups to evaluate their performance.

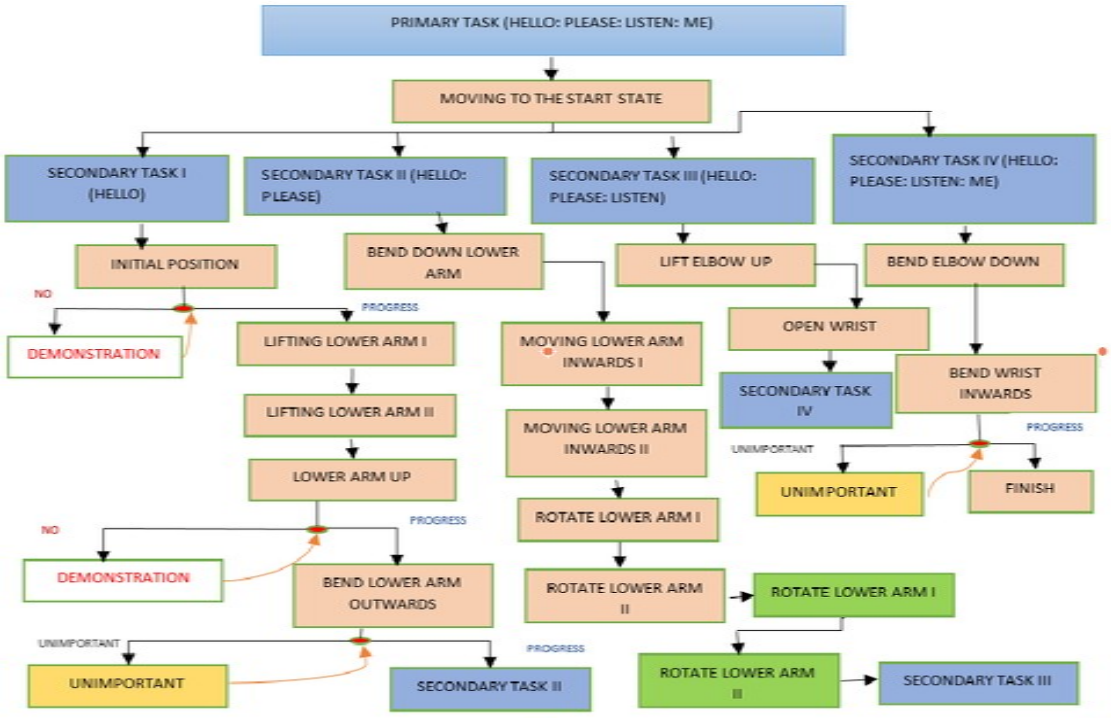

Figure 1. An example of a learned semantic hierarchical structure.

It is obtained by task segmentation, change point detection, and label and demonstration queries obtained in conversation with an expert demonstrator.

Phase-II

In this research, the humanoid robot Baxter motivates an individual extrinsically during the learning process using several positive reinforcers. During the interactions, the robot initially demonstrates several tasks to the participant, in ascending order of difficulty, involving the assembly of augmented-reality-tagged blocks into various patterns (Roy, Kieson, Abramson, & Crick 2018, March). If a candidate is successful in every task, external motivation has negligible effect in changing anything, because people feel comfortable and they are in a familiar situation. Thus, people who are happier will sometimes be less motivated to push themselves toward action compared to someone in a negative mood, who will be more motivated to exert effort to change their unpleasant state. Hence a negatively-valanced mood can increase, and positively-valanced mood can reduce, perseverance with difficult tasks. This may be because people are less motivated to exert effort when they are already satisfied with their performance. Frustration, in turn, may increase perseverance as people see greater potential benefits of making an effort. Thus if people in a negative mood get some positive reinforcer to overcome their challenge, their learning rate is expected to increase (Wong & Csikszentmihalyi 1991). Block diagram showing the human-robot interaction. In this research, the humanoid robot Baxter uses a reinforcement learning strategy to understand the effect of its reinforcement presentation on its human subjects, attempting to increase their performance over time. Here the subject pool is divided into sets of participants who receive no reinforcements, random reinforcements, or learned reinforcements respectively during their task performance. We compared the number of people committing more than three mistakes in each group, because we expect our reinforcement strategies to be more effective for subjects who are performing somewhat poorly. We also look at the overall number of mistakes committed by each subject group. We discovered that

participants in the learned group were more likely to perform well and committed comparatively fewer mistakes with respect to the other experimental conditions. We also learned that the robot's regret strongly correlates with the probability that a test subject makes more versus fewer mistakes.

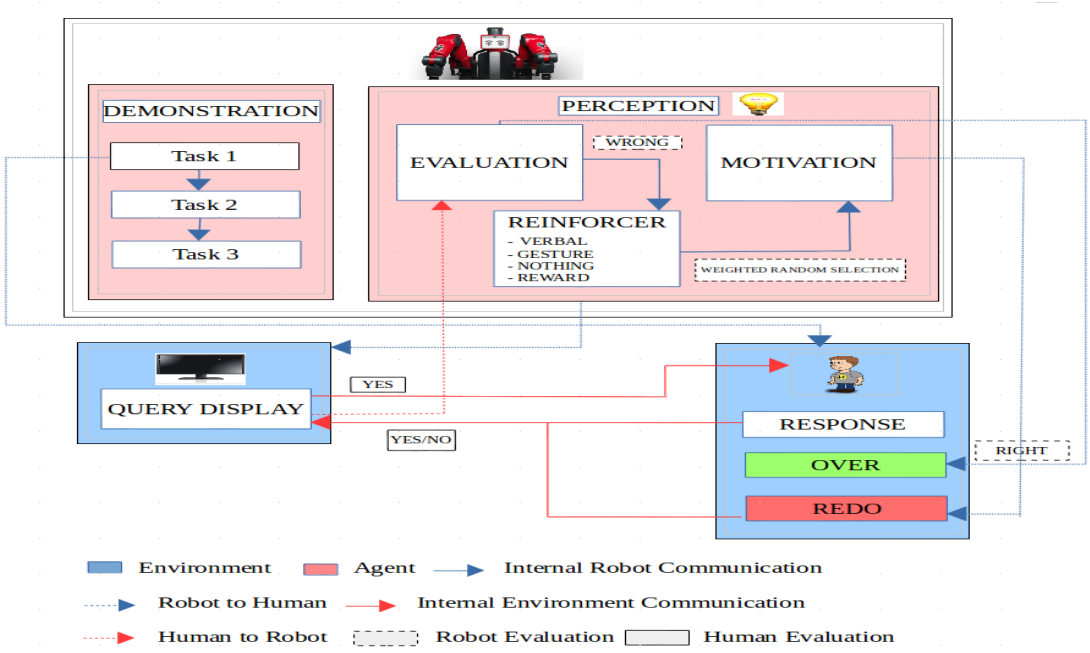

Figure 2. Block diagram showing mutual reinforcement learning.

Phase-III

In this research, the humanoid robot Baxter and a computer system (depending on the experiment) motivate an individual extrinsically during the learning process using several positive reinforcers. During the interactions, the robot or the system initially demonstrates the task to the participant, and then provides learned reinforcers to make sure the skill is transferring properly, using Simpson's psychometric model (Simpson 1972), and concurrently learns about their cognitive models. The expert uses a reinforcement learning strategy to understand the effect of its reinforcement presentation on its human subjects, attempting to increase their performance over time

(Roy, Crick, Kieson, & Abramson, 2019, Roy, Crick, Kieson, & Abramson, 2018). To identify the success of MRL-guided skill transfer, we divided the subject population into three groups where participants get no reinforcement, random reinforcement, or individually-tailored learned reinforcement (MRL) respectively. We compared the number of mistakes in each group, because we expect MRL to be more effective for subjects who are performing somewhat poorly. We discovered that participants in the learned group were more likely to perform well and committed comparatively fewer mistakes with respect to the other experimental conditions ($p < 0.05$ for Baxter) and ($p < 0.05$ in random group for the computer-based Tetris skill learning). We also determined information gain over time and how a machine's regret strongly correlates with the probability that a test subject makes more versus fewer mistakes. In addition, we produced confusion matrices demonstrating the effectiveness of MRL in the experiments using 5-point Likert scale data.

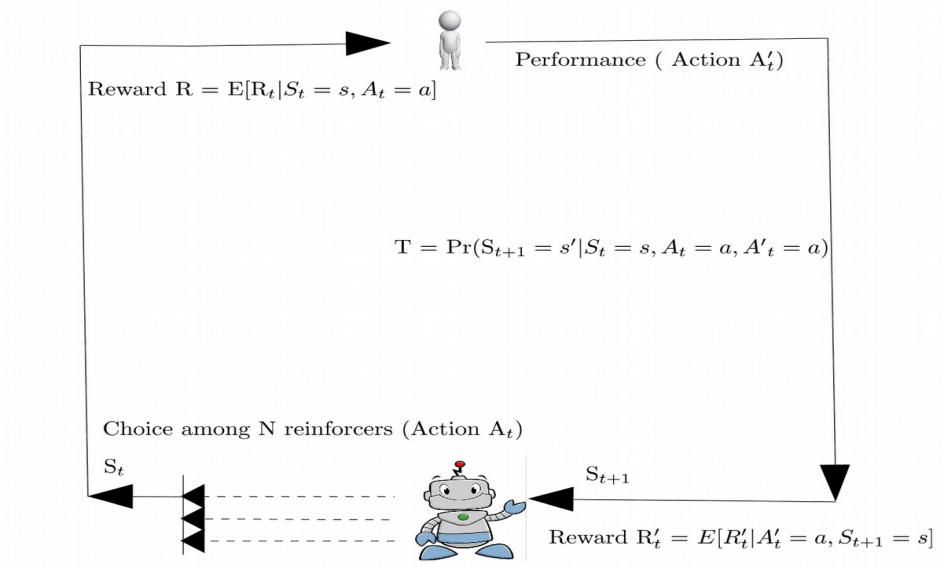

Figure 3. Mutual Reinforcement Learning.

CHAPTER II

RELATED WORK

Many contemporary researchers are working on identifying appropriate reward channels in human-robot collaborative frameworks to maximize performance. The following section briefly discusses this work along with the influence of positive reinforcer on motivating humans in skill transfer procedures.

Robot learning from demonstration

Konidaris (Konidaris, Kuindersma, Grupen, & Barto 2012) describes robots that can learn from trajectory demonstrations by constructing skill trees (CST). Chains from multiple human expert demonstrations can be merged into a single skill tree with a policy learning algorithm which efficiently increases robot learning rate. Crick (Crick, Osentoski, Jay, & Jenkins, 2011) and Knox (Knox, Breazeal,& Stone, 2013) illustrate that human experts directly teaching a robot is usually a better option than the same robot learning virtually, as humans have better understanding of the environment and a better decision making ability than unsophisticated robotic controllers. Similar work has been performed by Grizou (Grizou, Lopes, & Oudeyer, 2013) , where unknown human teaching instructions are utilized by the robot to improve its learning. This work addresses how a robot can use unfamiliar and noisy teaching instructions to

acquire knowledge to generate new tasks, and use that knowledge to improve its learning policy in an inverse reinforcement learning domain. While interacting, the robot tends to ask different questions to the end-user. Cakmak (Cakmak, & Thomaz, 2012 ; Cakmak, Chao & Thomaz, 2010)describes different platforms where the robot is trained to ask good questions and how their performance is improved via human feedback.

Robots Learning to Teach

Many scientists have started exploring this new area of robotics, where along with robot teaching we can gain other useful information about robot and human behaviors from their interaction. Spaulding (Spaulding, Chen, Ali, Kulinski, Cynthia Breazeal, 2018) introduced an integrated system for autonomously analyzing and assessing children's speech and pronunciation in the context of an interactive word game between a social robot and a child. This approach used Gaussian Process Regression (GPR), augmented with an active Learning protocol that informed the robot's behavior. Scassellati (Ramachandran, & Scassellati, 2015 ; Litoiu, & Scassellati, 2015) and Park (Park, Rosenberg-Kima, Rosenberg, Gordon, & Breazeal, 2017) have presented feedback-based human-robot interaction, demonstrating that if humans are guided by a robot at an interpersonal level, it increases the robot's perceived social reliability and makes humans more eager to interact with it. A robot learning from human feedback tends develop a mental model (Lee, Lau, Kiesler, & Chiu, 2005, April ; Scassellati, 2001) of its own which positively influences human cognition. Fasola et al. (Fasola, & Matarić, 2013) used socially assistive robots (SAR) to train elderly humans in physical fitness by motivating them. Yin et al. (Yin, Billard, & Paiva, 2015) described intelligent robot systems acquiring human-like writing style and then exploiting it to teach children. Fan (Fan, Tian, Qin, Li, & Liu, 2018) used neural network models to evaluate teaching strategies when one intelligent system is trying to teach another. Leite (Leite, Castellano, Pereira, Martinho, & Paiva, 2012) used robots to socially support children in a game scenario.

The robot not only increased the performance accuracy of the human learner, but also connected with them emotionally and provided social assistance throughout their learning process. This social support helped the children to build their self esteem and encouraged them to perform better. Humans have distinct teaching strategies (Khan, Mutlu, & Zhu, 2011) which can be effectively utilized in human-robot communication to build effective robot learners. A robot learning from human feedback tends develop a mental model of its own which can be later utilized to teach novice human operators. Scassellati (Strohkorb & Scassellati, 2016 ; Scassellati, Admoni, & Matarić, 2012) discusses human-robot collaboration for social good. If robots and humans can interaction an interpersonal level, achieving complex tasks is easier. In this work, feedback-based human-robot interaction demonstrates that if humans are guided by the robot or vice-versa, relevant questions are addressed, and with continuous collaboration the task becomes easier. If robots are able to teach human novice operators, this can improve their social reliability and enhance people's eagerness to interact with them. In this work, we describe human-robot interaction where the robot acts a teacher to guide humans to achieve complex sets of tasks. Cakmak (Cakmak, DePalma, Thomaz, & Arriaga, 2009) demonstrated how social learning strategies vary with the particular environment when robots are allowed to explore and learn from their surroundings. In this work, along with the effectiveness of MRL, we are also concerned with the idea of robots learning to be good teachers. We use a robot's own predicted regret and confusion matrices to evaluate its own cognitive model.

Reinforcement Learning Techniques to Identify Better Reward Channels

Rewards play a crucial role in both identifying and shaping a person's behavior. They not only tell us about a person's personality, but provide an influence channel when used effectively. Hence, recently many scientists are interested in researching appropriate reward channels that might increase task performance. Lopes (Clement, Roy, Oudeyer, & Lopes, 2013)

upgraded Multi-Arm Bandit techniques using different motivational resources to maximize skills and learning activities. He also researched the recovery of reward functions $R_{xa}(p)$ from expert demonstrated policies π_i (Lopes, Melo, & Montesano, 2009) to ensure active learning. The modification of traditional reinforcement learning algorithms using reward shaping produced important insights into how skill and accuracy can be improved for a particular task. Cooperative inverse reinforcement learning (CIRL) uses a human reward function $R : S \times A^H \times A^R \times \Theta \dashrightarrow R$ that maps world states, joint actions, and reward parameters to real numbers to establish useful human robot collaboration, where the robot is unaware of the initial reward. CIRL can be used in various platforms like active teaching, active learning, and communicative actions that are more effective in achieving value alignment (Hadfield-Menell, Russell, Abbeel, & Dragan, 2016). These researchers are also using multi-arm bandit techniques (MAB) to address the problems with assistive agents who can help human participants to select appropriate channels to maximize the cumulative reward (Chan, Hadfield-Menell, Srinivasa, & Dragan, 2019). Here the human does not know the reward function but can learn it through several interactions, whereas the robot only observes the human interactions and not the reward associated with it. Tabrez et al. used their Reward Augmentation and Repair through Explanation (RARE) framework for estimating task understanding where the autonomous agent detects potential causes of system failures and uses human-interpretable feedback for model correction (Tabrez, Agrawal, & Hayes, 2019). Nikolaidas et al. described a human-robot cross-training framework using reinforcement learning techniques where humans and robots switch roles to improve the overall performance (Nikolaidis & Shah, 2013) . Li et al. used MRL in automatic poetry generation using two models (local and global) which have some predefined criteria as rewards, and they learn from each other to pursue higher scores (Yi, Sun, Li, & Li, 2018). Griffith (Griffith, Subramanian, Scholz, Isbell, & Thomaz, 2013) discussed novel policy shaping algorithms and how motivations and reward signals can be used as a channel to impact human-robot partnerships in an HRI setting, simultaneously improving the future learning process of both humans and robots. Knox et al.

(Knox, & Stone, 2008 ; Knox, & Stone, 2009) designed a novel framework named TAMER which allows a human to train a learning agent to perform a complex tasks over continuous interaction. In our previous papers (Roy, Kieson, Abramson, & Crick 2018, March) we have also discussed how the robot updates its own cognitive model with each human interaction, improving the overall task performance through exploration-exploitation strategies. In this work we are not only extending the goal of our previous work, but also updating the MRL technique for better results.

Empathy and Positive Reinforcers

Empathy is based in the social-cognitive and behavioral ability to vicariously experience another person or animal's affect, and is critical in the social interactions of humans and some animals (Keskin, 2014 ; Lockwood, 2016) . Empathy plays a vital role in social interaction in all stages of human life and many contemporary researchers are working on empathetic robots that are designed to respond to human behavior and emotion with appropriate social cues. Empathy and adaptation may not be enough, however, since social responses are only one component of effective human-robot interactions. Instead, robot interactions that facilitate mutual learning with the human counterpart may prove more effective in a teaching environment due to the ability to learn, adapt, and create reinforcement feedback tailored to the individual.

Hence the ability to empathize has also been found to be a critical characteristic of effective teachers. In one study, teachers demonstrating more empathy were able to adapt the structure, behavior, and manifestation of empathy based on the group or individual and provide more effective teaching strategies (Mihaela, 2013). In our research, we used several positive reinforcers as reward channels to interact with the human participant. In order to effectively study human-robot interactions and learning, scientists have incorporated other socially-inspired tools

in addition to empathy. For example, auditory and visual cues are important in learning exchanges between humans and robots, especially when learning through demonstration (Koenig, Takayama, & Matarić, 2010). Furthermore, modeling demonstration learning using robots and humans has shown to be effective and the closer the demonstration technique was to typical social learning, the more rapport the participant felt with the robot and the more he or she learned (Sauppé, & Mutlu, 2015). Humans also demonstrate a need to share intentions with their social partners, and in order to mimic this with robots, the robot partner needs to mimic the social skills necessary to interact with humans and demonstrate shared intention (Dominey, & Warneken, 2011). The robot, in this case, demonstrated the ability to learn a goal and intentional actions linked to the goal through cooperative learning (Dominey, & Warneken, 2011). In these cases, behavioral interactions and social acceptance are critical components to the human-robot interaction. It is possible for humans to respond to perceived empathy from robot and computer interactions. Research shows that individuals perceive empathy through digital devices and computer-mediated interactions, and additional studies are developing robot-human interactions that more closely mimic human-human interactions using touch and visual interactions (Powell, & Roberts, 2017). Furthermore, empathy increases rapport between humans and robots, which is important for user comfort (Leite, Pereira, Mascarenhas, Martinho, Prada, & Paiva, 2013). This suggests that, while empathy is important for contextual comfort, it may not be the only component of a learning environment and does not indicate a human response for the robot. While scientists may have developed robots to mimic empathy that can be detected by participants, humans have yet to respond with equal attachment or empathy towards robots (Konok, Korcsok, Miklósi, & Gácsi, 2018). In other words, while adaptive empathetic robots may build some rapport with humans, the communication is only from the robot to the human; the robot is not necessarily responding in ways that may be necessary for human learning. A few researchers have explored various areas where positive reinforcement from robots had a large impact on children. Boccanfuso et al. (Boccanfuso, Barney, Foster, Ahn, Chawarska, Scassellati,

& Shic, 2016, March) investigated the difference in responses between children with or without autism with an emotion-stimulating robot using positive reinforcement in an interactive environment. Nunez et al. (Nunez, Matsuda, Hirokawa, & Suzuki, 2015) described the use of positive reinforcers to overcome the underlying challenges in motivating a child to continue learning and to share the experience with others. Kim (Kim, Berkovits, Bernier, Leyzberg, Shic, Paul, & Scassellati, B. 2013) addressed the unique positive effects and advantages a robot can have on autistic children, exploring areas where robots play an important role in the lives of specific individuals. We wish to investigate how a robot can develop an understanding of the underlying motivations and cognitive traits of individual people, so that it can shape its teaching strategies appropriately and enhance the learning process.

Positive Reinforcers in MRL

Mutual feedback between robot and human has become increasingly important in human-robot interactions. Interactivism (Bickhard, 2009) and process-oriented robots have been challenging in the past since there is a necessary balance between environmental stimuli and feedback and the adaptation of software and processes that can adapt and change with them (Stojanov, Trajkovski, & Kulakov, 2006). Robots using socially-inspired reinforcement including verbal and behavioral feedback have only shown modest results, and studies have suggested that a more targeted approach tailored to the individual would be better suited for future robot-human interactions (Ferreira & Lefevre, 2015). To better facilitate natural social interactions and engage with the learning environments of humans, robots need to adapt and respond appropriately to each individual. Positive reinforcement increases learning in all animals and promotes voluntary behaviors of animals, but the reinforcement tools need to be species-specific and based on individual preferences and experiences. In this sense, if robot-human interactions are to use socially-derived reinforcements as teaching tools, researchers need to take into account not just

human social interactions, but individual differences as well. This means that the robots need to be programed with an understanding of individual-specific approaches to interactions based on principles of learning and be able to adapt and respond in ways that are tailored based on the individual's unique responses. The scientists in this study have developed a novel approach using mutual reinforcement learning where both the robot and human act as individual empathizers who can act as reinforcement learning agents to achieve a particular task. Thus in this paper the humanoid robot Baxter and a personal computer not only adapts or empathizes with its human participant but also takes a step forward to encourage them and achieve their goal.

Exploration Exploitation Strategy

The exploration/exploitation dilemma (Audibert, Munos, & Szepesvári, 2009 ; Baranes, & Oudeyer, 2009) is a common problem, where decision makers can either jump to a conclusion and make a decision on the basis of the partial knowledge they currently possess, or rather wait and invest more time and effort in accumulating further information, with the hope that a broader perspective will lead to a better decision in future. In our research, Baxter attempts to probe and understand a specific aspect of a human minds cognitive orientation toward particular reinforcement strategies, on the basis of this exploration and exploitation trade-off, where human performance acts as the reward.

CHAPTER III

TECHNICAL DESCRIPTION

Mutual reinforcement learning (MRL) deals with the scenario where both humans and autonomous agents act as reinforcement learners for each other, identifying the path to achieve maximum reward. In this instance, the robot initially acts as an expert and its human counterpart as a novice. In MRL, one agent's action is mapped as a reward to another. Here the agents, as they are unaware of each other's incoming actions, discover the appropriate reward channel over continuous communications with one another. The autonomous agent acting as an expert learns about the appropriate reward channel through an exploration-exploitation tradeoff (Audibert, Munos, & Szepesvári, 2009). The action of the novice agent (judged in terms of the agent's performance) not only affects the immediate rewards but also the expert's next action. The expert does not immediately jump to a conclusion about the decision to be made, but rather invests more time and effort in accumulating further information, with the hope that a broader perspective will lead to a better decision in the future. On the other hand, the humans interpret the actions of the robot (or computer) agent as a reward, which influences their performance in learning the task. In this paper, MRL is implemented in a skill transfer scenario, where the autonomous agent is trying to teach a human some complex task, while updating its own mental model (Sutton & Barto 2018) at the same time.

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 € A or A' in state s, and r € R and R' : (S, A) or (S, A') ⇒ R and R' are the reward functions. Since, in MRL, the action of an agent is the reward to another and vice versa, the tuple can be simplified as follows: Novice={S, A' , T, R}, Expert={S, A, T, R'} where if the novice 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. Here the reward for the novice r € (r_1 , r_2 , r_3 ...) is selected by an exploration-exploitation tradeoff where r_1 , r_2 , r_3 are all different kinds of reinforcers mentioned in Chapter 4. In the case of MRL, we have a verbal, hint, gesture and simple feedback for the robot whereas there are seven different reinforcers in case of Tetris. Therefore the expected rewards in both the cases in a state action pair can be written as a two-argument

function r : SXA ⇒ r€R and r : A'XS ⇒ r€R' .

$$r(s, a) = E[R_t|S_t = s, A_t = a] \quad (3.1)$$

$$r(s, a) = E[R'_t|A'_t = a, S_{t+1} = s] \quad (3.2)$$

The above equations imply that whenever a novice makes a mistake at time t, the robot takes an action a_t in that state s_t to positively reinforce the participant, who on the other hand takes an action a'_t and rectifies the mistake moving to the next state. Here the state is the pattern-making task in the case of Baxter, while in the case of Tetris, the players are asked to restart the game rectifying the mistake. Hence the reinforcement learning agents give rise to a sequence or trajectory that looks like the following if the novice keeps on making a mistake: S_t , $A_t(R_t)$, A'_t

(R'_t), S_{t+1} , A_{t+1} (R_{t+1}), $A'_{t+1}(R'_{t+1})$, S_{t+2} . The above sequence denotes the condition if a participant keeps on making a mistake at one point. The sequence will stop with the correct action of the novice learner. However, the robot keeps on evaluating the other sections and if a mistake occurs again, the same behavior is repeated.

Therefore the state transition probability T :

$$p(s' |s, a) = Pr(S_{t+1} = s'|S_t = s, A_t = a, A'_t = a) \quad (3.3)$$

Robotic mutual reinforcement is based on psychological principles of social reinforcement and inclusion and is intended to improve skill transfer by adapting to the reward value systems of an individual. In order to effectively teach a skill, the instructor relies on the principles of learning theory and basic operant conditioning and positive reinforcement. Reinforcement, whether through the addition of a reward (positive reinforcement) or the removal of something aversive (negative reinforcement) refers to techniques used by trainers and instructors whose goal is to increase the likelihood of the behavior being repeated.

Behavioral psychologists rely on the principles of positive reinforcement as the primary means through which to teach and shape behaviors in both humans and animals. This type of shaping is considered ideal since the individual participant or subject is rewarded for the correct behavior and associates a specific behavior with a specific reward. The subject is therefore more likely to repeat the desired behavior in the future (as opposed to negative reinforcement when the subject is trialing behaviors to avoid an aversive stimulus). In the case of mutual reinforcement, the robot engages in activities and behaviors that positively reinforce the correct behavior of the participant. Learning theory dictates, however, that the value of rewards in positive reinforcement are subjective in nature and highly dependent upon the individual. This means that a reward (R) may be of high value to one individual and of no value to another. For adequate learning to occur, the reward must therefore be tailored to each individual. In the case of mutual reinforcement, the

robot is equipped with a range of rewards that might be of value to the subject. Given that all the subjects are human, the programmed rewards are tailored to humans and were designed based on the species, culture, and potential individual differences of the target population. Humans are social and, in performance environments, are highly responsive to social inclusion and exclusion (Cheung, & Gardner, 2015), making social signals of high value to almost every human. The social rewards are designed to mimic culturally-appropriate interactions to which the participants are accustomed, which promotes comfort and relaxation and builds rapport. The ability to adapt and read the human through behavioral feedback establishes a baseline of communication and language that is novel and unique between the robot and the human, effectively mimicking normal social relationships and social learning paradigms. When a autonomous machine is equipped with various means of social reinforcement in combination with algorithms that allow for adaptations to individuals, there is a greater chance of finding a reward that is of high subjective value for each participant. This can only be done if the machine is first equipped with these tools and then programmed in a way that allows it to perceive, adapt, and adjust rewardsbased on the feedback from participant progress. Mutual reinforcement is therefore a promising approach to skill transfer between a robot and human.

Initially, the robot assigns a uniform prior across its potential reinforcement behaviors, and they begin with equal weights. When a subject is given a particular reinforcement, the robot evaluates her performance on the immediately following subtask, and reweights its reinforcement strategy appropriately.

$$S_t = \{\nu\varphi s^{+}_{t-1}\,,\ \nu\,(1-\varphi)s_{t-1}/(|S|-1)\ \forall s \in S_{s\ \neq s+}\} \tag{3.4}$$

where S t is the weight distribution over all reinforcement strategies at time t, s^{+}_{t-1} is the particular reinforcement strategy chosen at time t − 1, φ is 1 if the subject successfully completed the subtask immediately following the previous reinforcer, and 0 otherwise, and ν is a

learning rate parameter. In our experiments, ν was taken to equal 0.03. Thus after several iterations, the robot can conclude which particular reinforcements are inducing the candidate to perform well; in other words, after receiving particular reinforcers, the candidate does not underperform. The two experiments use the reinforcers differently. In the first experiment, reinforcements are provided when the human is performing well, but the results of that experiment led us to the robot providing reinforcement when a candidate was disappointed by his performance. The following sections illustrate Algorithm 1 and the notations associated with it. It also explains the choice of selection of certain parameters and their impact in the experiment.

Optimization of Reinforcers

The above algorithm is implemented in both robot and computer gaming platforms inducing MRL. This method is directly applied to the problem of searching for an appropriate reward channel preferred by individual human participants during the skill transfer task. Here each reinforcer is influenced by the participant's cognition and performance, and this evaluation directs which reinforcer will be considered next. Hence, in this method, the robot and the human are successively generating and evaluating attempts to obtain incremental improvements for each other.

Algorithm 1: Mutual Reinforcement Learning Algorithm 1

Input: $V_n \leftarrow \{set\ of\ items\ with\ uniform\ weight\ distribution\}$
Output: $V_{n_{updated}} \leftarrow \{set\ of\ items\ with\ updated\ weights\}$
$V_{n_{ith}}$ *is given out on the on the basis of weighted random selection*
while $V_{n_{updated}}, (mistakes)$ **do**
  if $V_{n_{ith}} \leftarrow success$ **then**
    $V_{n_{ith}} = V_{n_{ith}} + \alpha;$
    $V_{n-n_{ith}} = V_{n-n_{ith}} - (n-1)/\alpha;$
  else
    $V_{n_{ith}} = V_{n_{ith}} - \alpha;$
    $V_{n-n_{ith}} = V_{n-n_{ith}} + (n-1)/\alpha;$
  end
  $EWMA(V_n) = (V_{n_{updated}} * \phi) + (Previous\ EWMA(V_n) * (1-\phi))$
  $V_{n_{updated}} = EWMA(V_n) + \sigma(V_n)$
  $V_n = V_{n_{updated}}/(V_{n_{updated}}.sum)$
end

Figure 4. Algorithm 1: Mutual Reinforcement Learning Algorithm 1

Figure 3. refers to the MRL concept where the participant rectifies the mistake after getting a positive reinforcer. In the particular studies reported here, Baxter has four distinct reinforcers at its disposal, whereas for Tetris there are seven. V n denotes the weight vector assigned to n reinforcers, which are initially a uniform probability distribution summing to 1. The reinforcers are given out on the basis of weighted random selection to meet the exploration-exploitation criteria, but since all the reinforcers are uniformly weighted at the beginning,weighted reinforcement in the first step is of no significance. If the reinforcer Vn_ith given out from the set is a success then Vn_ith + α. Here i represents the particular reinforcer thatis provided by the robot or the machine to motivate people. α is a small positive fraction called the step-size parameter, which influences the rate of learning. The value of α is chosen empirically based on the observed performance of exploration and exploitation. Figure 5. and Figure 6 denotes the entropy fluctuation of the system with the suitable alpha value. Selection criteria for the value of α are discussed in detail in the next subsection. The value of α adds to the present weight of the reinforcer denoting its success. (n − 1) denotes the number of remaining reinforcement strategies, and the value of α is equally distributed among them. If V n_ith is successful, α is added to it and (n−1)/α is subtracted from all the rest of the reinforcers. In contrast, if Vn_ith is a failure, then α is subtracted from it and (n−1)/α is added to the rest of the reinforcer. After this step the weights of V_n are updated. The mutual information shared among them is obtained from several interactions and the values of the reinforcers get updated every time with probabilities associated with faster skill transfer. The robot tries to learn about the person's behavior and performance level and then applies this knowledge to motivate the individual. We used an exponential weighted movingaverage (EWMA) to gain information about the most recent interactions. For Baxter, since the number of reinforcers is fewer, we used the past three interactions and for Tetris we used five. EWMA only provides information about recent interactions, but we need to understand the variability of the reinforcers' success over a longer term. Hence we maintain the value of two standard deviations σ(V_n) over the EMWA

values in order to better notice and interpret success. Then again the probabilities of all the reinforcers in the set are updated and prepared for the next interaction. The robot stops giving out the reinforcers when the participant stops making mistakes. MRL, an implementation of which we have described in the above algorithm, is a novel concept in the field of traditional reinforcement learning and can be implemented in several algorithmic approaches to get significant results.

Choice of Parameters

In Algorithm 1 we used several parameters whose values are tailored depending on the experimental requirements. The parameter α is chosen on an empirical basis. We used numbers ranging from 0.01 to 0.07 in case of Tetris and 0.01 to 0.04 in case of Baxter. The range of numbers are selected on the basis of the number of reinforcers used in the experiment. We calculated entropy $\hat{H}$ depending on the weight of α to determine the mutual information gain over interactions and whether the autonomous agent is optimally exploring and exploiting. Using the maximum entropy principle, we know (Shannon, 1948) that entropy reduces over time with the information gain. We calculated the entropy using values chosen from the ranges given above to determine the most suitable α value.

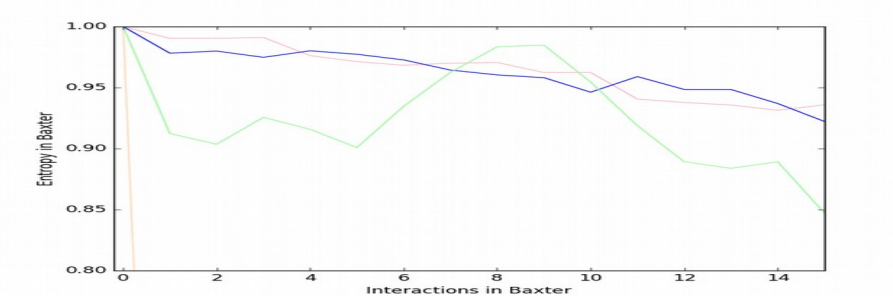

Figure 5. Rates of entropy decrease in Baxter for different α values.

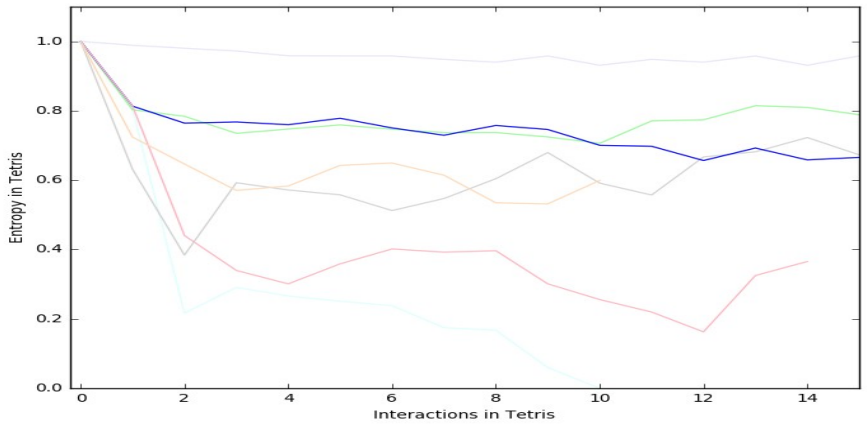

Figure 6. Rates of entropy decrease in Tetris for different α values.

The lines in lighter color show different α values we considered for Baxter and Tetris and and the dark blue line exhibits the one we selected for our experiments, namely 0.015 for Baxter and 0.05 for Tetris. These values not only gain information linearly over time but also trade off satisfactorily between exploration and exploitation. In the process of information gain, the rate of decrease of entropy is not always perfectly linear; all the graphs of different α values are accompanied by spikes due to exploration-exploitation tradeoffs. We designed the above algorithm not to behave greedily because exploring different reinforcers is necessary while breaking the monotony of the task. From these figures we can see that at the beginning of the experiment the entropy is maximum for all α values and gradually decreases over interactions.

The value of φ is selected in such a fashion that the robot considers the last three interactions of Baxter and five for Tetris. φ is denoted as the multiplier. In our previous work (Roy, Crick, Kieson, & Abramson, 2018 , we used the robot's experience from the beginning to the task for the reinforcer selection and found out that sometimes people prefer more than one reinforcer. We theorized that their preference might have changed over interactions, and hence focused on recent interactions for better performance.

MRL and cognitive models

In traditional reinforcement learning designing an appropriate reward signal is a critical part of the application process. Various researchers have coined novel techniques to solve this issue (Abbeel, & Ng, 2004). In contrast, processes like inverse reinforcement learning (IRL) learn from a expert's behavior, where an agent tries to infer the reward signal to achieve a particular goal. In neither of these cases does any two-way interaction between the expert and the agent take place, and therefore they do not gain the advantage of situational feedback which is important 19during a learning process. To achieve a particular task, both expert and novice both should exchange feedback through appropriate reward channels. In practice, designing appropriate reward signals is often an informal trial-and-error search for a reward signal that produces acceptable results. In MRL, the expert explores and exploits the reward signals in the course of judging the novice's actions and performance while accomplishing the task. Hence if the novice learns slowly, fails to lean or learns incorrectly the expert cooperates to improve the student's learning during the process. This is a sophisticated way to find good reward signals, since feedback is given while accomplishing a subgoal and the expert can slowly guide the agent towards the overall goal. Hence unlike other reinforcement learning strategies, MRL is a complete model that supports task learning with human-robot interaction simultaneously learning about the reward preferences. In MRL, since the expert cooperates with the novice during the learning process, it also becomes aware of the cognitive models involved, which in turn leads to the design of better reward signals. To explore the efficiency of the process, we calculated the machine's regret and the mutual information shared between the agents using Shannon's entropy $\hat{H}$. Regret is defined as the difference between the reinforcer with maximum weight and the reinforcement strategy selected, i.e. $R = s_{max} - s + .$

Property 1: **In MRL, an autonomous agent fails to identify the cognitive orientation of a participant if it crosses k time steps as after k steps no change in entropy occurs, which means no information gain.**

The above matrix M is a transition matrix with state space S, where |S| = X is possibly infinite. N denotes the number of reinforcers used and α and β are the weights that are added to the system depending on the exploration-exploitation tradeoff. The above matrix is prepared on the basis of Algorithm 1. Now let π^T be a row vector denoting a probability distribution on S: so each element π_i denotes the probability of being in state i,

$$M = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \dots \\ \frac{1}{N}+\alpha & \frac{1}{N}-\beta & \frac{1}{N}-\beta & \frac{1}{N}-\beta & \frac{1}{N}-\beta & \dots \\ \frac{1}{N}+\alpha-\beta & \frac{1}{N}-\beta-\beta & \frac{1}{N}-\beta-\beta & \frac{1}{N}-\beta+\alpha & \frac{1}{N}-\beta+\beta & \dots \\ . & & & & & \\ . & & & & & \\ . & & & & & \end{bmatrix}$$

and $\Sigma^{X}_{i=1} \pi_i = 1$, where $\pi_i \geq 0$ for all i = 1, . . . , X. The probability distribution π^T is an equilibrium distribution for the above matrix if $\pi^T M = \pi^T$. That is, π^T is an equilibrium distribution if $(\pi^T M)_j = \Sigma^{X}_{i=1} \pi_i p_{ij} = \pi_j$ for all j = 1 . . . X

That is, π^T , π^{T+1} will have the same values and so on. This is because the values achieve numeric stability after π^T . That means M^t converges to a fixed matrix with all rows equal as t → ∞. At this point, no further change in the Shannon entropy $\hat{H}$ for π^T , π^{T+1} will be observed. Entropy in a system denotes its information gain: a decrease in entropy means more information gain. Here π_1 has the maximum entropy 1 which decreases linearly over time with information gain. Now if $\hat{H}$ does not change with time t, that means the robot is not gaining any further

information about the participant's cognitive orientation. Hence we can conclude that cognitive orientation of a participant can only be found ≤ k steps in the transition matrix.

Property 2: If MRL converges to stationary distribution over time (equilibrium), it is independent of the reinforcers used during the interaction.

From property 1, we determined the convergence criteria of M . Hence we know P that matrix M converges to π^T , $\Sigma^X_{i=1}\ \pi_i = 1$ for a large value of t. Now the stationary or equilibrium distribution can be found out by solving $\pi(M - I) = 0$, where I is the identity matrix. If a matrix M reaches equilibrium at $t \rightarrow \infty$, we know that the cognitive orientation of the candidate is undetermined. Hence we assume that when equilibrium is achieved all the reinforcers (r_1 , r_2 , r_3 , ...) are utilized and they failed to affect the human cognition. Hence we can conclude that if MRL converges to the equilibrium distribution, then it is independent of the reinforcers used during the interactions.

CHAPTER IV

REINFORCERS

Motivation and reinforcers

The human mind is a cognitive structure that consists of memory, decision making, perception, thoughts, emotions, and so on. These features act differently when they are influenced by external factors like stress or negativity on a regular basis. Thus, understanding the needs of a human mind under particular circumstances can be difficult (Wong & Csikszentmihalyi 1991), especially when those needs are dynamic or specific to a certain individual. In this research, Baxter attempts to identify the reward orientation of the particular human with which it is interacting, depending on the task performance. In this section, we first describe the robot's evaluation procedure used to assess the learning performance of its partner. We then describe how the robot tries to identify the best course of action to improve its own teaching performance. The exploration/exploitation dilemma (Audibert, Munos, & Szepesvári, 2009) is a common problem, where decision makers can either jump to a conclusion and make a decision on the basis of the partial knowledge they currently possess, or rather wait and invest more time and effort in accumulating further information, with the hope that a broader perspective will lead to a better decision in future.

In our research, Baxter attempts to probe and understand a specific aspect of a human mind's cognitive orientation toward particular reinforcement strategies, on the basis of this

exploration and exploitation trade-off, where human performance acts as the reward. The robot employs random selection among potential reinforcement behaviors, weighted by its current mental model of its human partner's motivation. When Baxter is trying to motivate an individual, it provides a positive reinforcer. Four different kinds of positive reinforcement are used in this process: verbal, reward, gesture and none . The autonomous agents weigh several positive reinforcers in this research to motivate the students if they commit any mistakes. In this section we discuss the reinforcers used by the robot and the computer during the experiment and the effectiveness of Simpson's psychometric model.

Reinforcers used with Baxter

When mistakes are made, Baxter forms a sad face and gives out the positive reinforcer to encourage participants (Fitter & Kuchenbecker, 2016), and when they perform correctly after the correction it forms a smiling face. Other than that, Baxter maintains a neutral face throughout the task. The following reinforcers are given out depending upon the subject's assigned experimental group.

Verbal reinforcer: When using this reinforcer, the robot asserts that it is trying to encourage the subject with some positive feedback. Since Baxter does not have its own audio interface, we used speakers to produce the robot voice. In our experiment, if the subject makes a mistake, the robot will verbalize something like, "Sorry dear, don't worry. You can do it".

Hint-based reinforcer: This takes the form of a hint given to the participant during a task. The hint does not provide the correct answer but tries to influence the subject's thought process so that it increases the learning rate of the participant. For example, during the pattern making process, if a candidate places an incorrect marker, Baxter suggests flipping the marker box and

trying the other side, before rejecting the block entirely. Thus people can track the blocks they have already tried to place in a particular spot.

Simple-feedback reinforcer: In this case, the robot only identifies the correct or the incorrect marker. It doesn't attempt to induce any kind of positivity or motivation in the participant. This is because some people are not fond of external motivations and only a rectification in the task can influence them to perform better.

Gesture-based reinforcer: In this case, the robot adds a consoling gesture by patting at the student's back and also provides positive verbal feedback as referenced above.

Reinforcers used with Tetris

In the case of Tetris, seven different positive reinforcers are used during the interactions. We increased the number of reinforcers because in a fast-moving gaming scenario, we anticipated more interactions per session. All the positive reinforcers are displayed in an audio-visual setting whenever a participant makes a mistake. Here all of the reinforcers provided some sort of hint for the player to perform better. For example, whenever a player is playing for too long without scoring any points, reinforcers are provided such as "Clear the lines quickly for faster score" or guiding the player to check for the upcoming blocks to plan the next move ahead. The type of incentives are manipulated according to the platforms we employed to demonstrate the effectiveness of MRL.

Simpson's psychometric model

In both of the above platforms, Simpson's psychometric model (Simpson, 1972 ; Simpson, 1966) is used to identify the skill transfer. This model characterizes the principles of skill

evaluation, which are closely linked with important aspects of human cognition. It is widely used by teachers, professional specialists and scientists to evaluate curricular problems with greater precision. Simpson's psychometry domain is defined in form of a taxonomy which gives us a clear idea about how knowledge is acquired by an individual and how that is later applied to execute tasks. Simpson's psychometric model is broadly classified into Perception, Set, Guided Response, Mechanism, Complex Overt Response, Adaptation and Origination. Perception is related to the awareness of the present situation. Set is the eagerness of the human participant to volunteer for the task. Guided Response is the early stage of learning a complex skill with the help of an instructor. After the participant has learned the task, the later stages of the psychometric domain involve applying the training. Mechanism is the immediate step to demonstrate basic proficiency with respect to a simple application. Complex Overt Response is associated with skillfully applying complex versions of the same task with greater proficiency.

Adaptation signifies complete learning, where individuals can respond to uncertain events, while Origination is the last phase of learning where humans can generate new ideas from their knowledge. Here the experiments are designed to confirm the feasibility of Simpson-based skill evaluation. Since the tasks are designed in a lab setting, we only used a few of the above categories to determine the skill transfer process. The following section discusses the

Figure 7. The mistake is rectified by the participant and Baxter forms a happy face

Figure 8. Participants make a mistake and Baxter forms a sad face while providing a positive reinforcer to encourage the participant

Figure 9. A reinforcer is provided to rectify the mistake. experimental models and the findings associated with them.

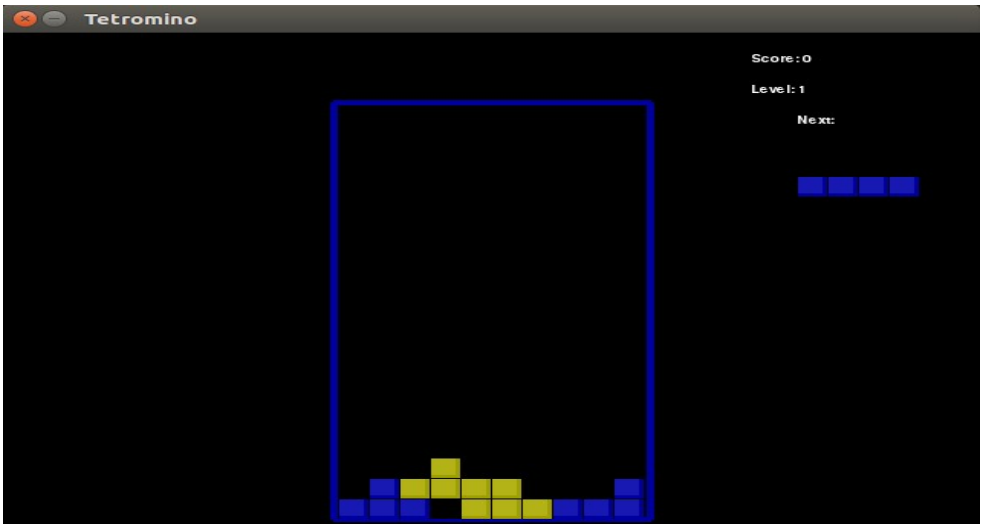

Figure 10. The participant has made a mistake in the Tetris game (allowed a gap to form in a line)

CHAPTER V

EXPERIMENTAL PROCEDURE

The following section describes the experiment conducted so far in learning and teaching domain.

Experimental design: robot teaching via demonstration I

This experiment involves performance analysis of participants when they are taught by the robot acting as an expert. The robot taught them the language motion ‘Hello, Please listen to me’ used for the experiment involving experts. In this experiment, the subjects are asked to imitate the same task they are taught by the robot, and return to demonstrate the task again after an interval of at least three days. The goal of the experiment was to evaluate the teaching action of a robot while interacting with a novice. If the novice human operator successfully manages to imitate the task taught by the robot, then we can infer that the robot has taught that person well, especially if the skill persists over time. Since semantics play a great role in providing useful information, we used our learned semantic model for teaching with semantic labels along with the corresponding gesture. The experiment is a between-subjects study where one group receives the benefit of semantic structure, while the other only receives demonstrations.

Procedure

In this experiment n = 38 subjects were involved. There were 20 people in the semantic labeling group and 18 in the control group. Participants in these groups learned to control the robots arm using a joystick controller to produce the ‘Hello, please’ sign language phrase which the robot had learned from expert demonstrators. The self-reported joystick proficiencies of the participants were noted at the beginning of the experiment and used as a control. The robot performs a motion and the participant attempts to duplicate that motion using the controller device. During the experiment the participants were provided with necessary information regarding the robot and the functionality of the joystick controller. They were allowed to take notes for their convenience. On the first day of the experiment, both groups of participants were allowed to see Baxter demonstrating the task as many times as they wished, and could practice for half an hour to get acquainted with the robot. Since this group of participants were intended to be novice human operators, they were not given any human guidance from the researchers and were only allowed to learn from the robot. Subjects in the semantic structure group were ‘taught’ using semantic labels assigned to each movement, which were previously developed through active engagement with an expert. These labels were broken down to indicate the smaller actions that make up the entire task. Participants followed the robots instructions to learn the movements necessary for using the joystick controller to perform the same actions, thereby mimicking the motions of ‘expert’. The participants in this group can see the labels and task structure on the monitor during the task and are also given out handouts containing necessary information about Baxter. During the experiment, since Baxter is teaching them a new motion, no other human guidance is involved. Participants in the no semantic structure group were ‘taught’ without semantic labels, so that there are no associations between each smaller movement and any assigned categorization or word. Participants are expected to follow the robot’s demonstrations to

learn the movements necessary for using the joystick controller to perform the sign language task. Participants in this group were only provided with the handout containing important information about Baxter and the joystick controller.

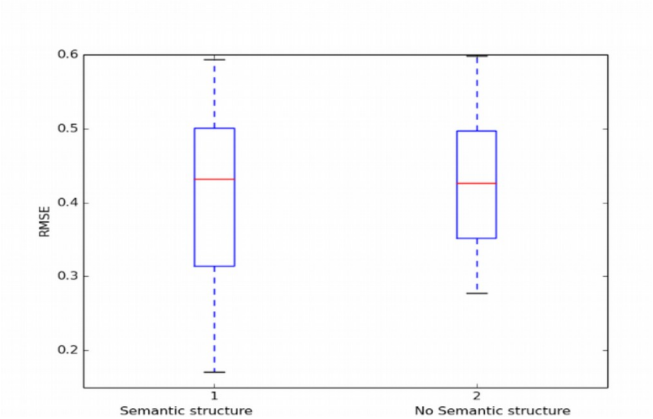

Figure 11. Performance analysis of the participants on the first day of the experiment.

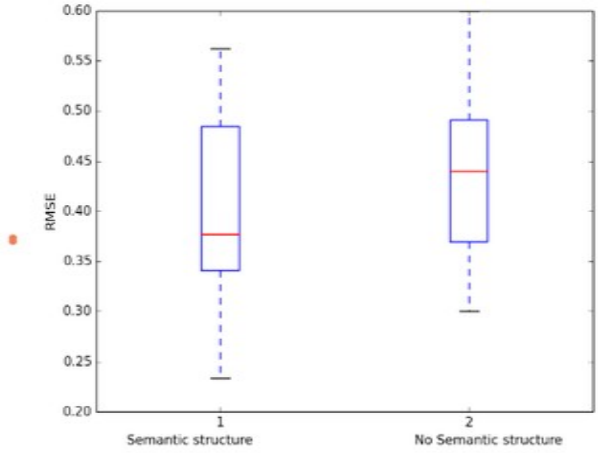

Figure 12. Performance analysis of the participants on the second day of experiment.

All of the participants in both groups were asked to return after at least two days and were asked to perform the same movements on the robot. They were only given one chance to perform the movement and were not allowed to practice or see any demonstration, although they

did first have the opportunity to practice random movements with the controller. The subjects in the no semantic structure group were asked to move certain joints to perform their motion, whereas the people in the semantic group were provided the semantic labels with which they were taught, as instructions to execute the corresponding motion. No significant difference between groups ($p \approx 0.5$) Difference is suggestive but not significant ($p = 0.18$) The above figures shows the RMSE values of both the group of participants scaled on the basis of their joystick proficiency.

Subjective performance evaluation

In Figure 11., which represents the initial training session where the robot demonstrates the task, there is almost no difference between the performances of the two groups. This is because the subjects were allowed to see the task demonstration as many times as they wished and to practice it several times to achieve adequate performance. Subjectively, we observed that subjects in the semantic structure condition asked for fewer demonstrations than those of the group without such structure. Since we determined that joystick proficiency played an important role in subject performance, and subjects were randomly assigned to the two groups, we found that the mean self-reported joystick proficiency of the non-semantic structure group (mean of 3.13 on a 5-point Likert scale) was significantly higher than that of the semantic structure group (2.51). Even when controlling for this, however, no significant difference in task performance was found. Figure 12. demonstrates that the subjects in the semantic structure group may have been able to retain their task expertise over a period of several days, compared to the subjects who were not provided with the same level of semantic assistance by the robot. A t-test shows a p-value of 0.18, suggestive but not a significant difference. Again, this data is after controlling for self-reported joystick proficiency. Subjects in the semantic structure condition may have had a

more thorough understanding of the task, but participants in both groups had access to their notes, taken during the initial teaching encounter, and this may have muted the effect. Even so, the non-semantic group did perform less well, and a larger fraction had larger errors than in the semantic group just not to the level of statistical significance.

Experimental Procedure II

n = 110 participants were recruited for the experiment (aged 18-20, 28 male, 82 female). The no-reinforcer group contains n = 35, the random reinforcer group contains n = 22 and the learned model group contains n = 53 participants. The blocks that are used in this experiment are two-faced having different markers on each side. During task 1, Baxter initially demonstrates seven markers, explaining their numbers and markings to the 29candidate. They are then randomly asked to identify two markers out of seven. A participant proceeds to the next task only if the first is finished successfully. Accordingly, the markers the candidates study in task 1 may not be repeated in task 2, as they are all shuffled before each task. The robot is only responsible for placing the markers in their respective positions. Since markers are shuffled randomly, each participant is given different patterns.

In task 2, subjects are asked to solve some general math questions as a distraction while the robot constructs the pattern, in order to reduce the available observation time for the participant. When the robot is finished making the pattern, they are asked to turn back and observe the pattern for 30 seconds. The blocks are shuffled again and the students are asked to recreate the pattern in 50 seconds. After each task, the robot inquires if they require more time. During task 3, the subjects are allowed to observe the pattern making process, but are not given any additional observation time. Baxter's pick-and-place manipulation is fairly slow, and it takes

almost a minute for it to create the larger pattern. Out of the three subject groups, substantially more participants in the learned reinforcement strategy group were able to advance to task 3.

Positive reinforcer on success

Initially an experiment was carried out where Baxter uses positive reinforcers as appreciation if the person performs well in a task. n = 19 participants are invited to the laboratory where they are allowed to interact with Baxter. However, the different groupsof people (none, random and learned) did not show any significant performance difference, judging from the number of mistakes they made in each condition. The median number of mistakes is the same across all conditions, although their overall range of mistakes varied somewhat. The subjects indicated in conversation with the researchers that when each participant is performing so well in all the tasks, they barely cared about the reinforcer from the robot because they are performing well anyway. Hence we redesigned the experiment in such a manner that the reinforcers can influence the subjects during their performance and the robot can help them better to accomplish the task.

Positive reinforcement on failure

In this experiment, the positive reinforcements are provided by the robot if the candidate has unsatisfactory performance at any point. Chapter 3, explains how the reinforcement strategies are adjusted according to the human orientation uncovered by the robot's exploration and exploitation of effective strategies. Here, the candidates are divided into three categories, where they receive no reinforcement at all, a random reinforcement, or a reinforcement selected according to Baxter's understanding of what motivates the particular individual. In the no-reinforcement group, Baxter only demonstrates the task and declares 'Right' or 'Wrong'

depending upon the performance of the participant. In the case of the random and learned model categories, the robot gives out positive reinforcers in the form of a reward, gesture, verbalization or just simply saying ‘You are right’. Baxter also changes its facial expression (Fitter & Kuchenbecker, 2016)on the basis of the candidate’s performance. Generally, Baxter puts up a neutral face while demonstrating the task, but if a candidate performs correctly, Baxter’s face turns green with a smile, while it makes a sad face and turns red when wrong. The facial expressions are also applied as a form of reinforcer. The goal of this experiment was to determine if, when the robot has lower regret, whether the learner makes correspondingly fewer mistakes with time.

Subjective performance evaluation

Figure 13 shows the number of mistakes made by each participant, used as a metric to evaluate performance. Out of all the participants who performed the experiment, there are some within each experimental group who did not make any mistakes. Those participants did not receive any reinforcements regardless of which group they were assigned to, so are not considered as a part of the mistake data. Also there are cases where the participants responded to more than one reinforcer or made so many mistakes that the robot could not determine their reward orientation. From within the learned model group, out of 53 participants 18.87 percent of people did not commit any mistakes, and the orientation of 35 percent of the participants could not be determined by the robot under the experimental conditions. This means that the robot successfully learned a good teaching strategy for slightly fewer than half of the participants.

Figure 13 considers those participants of the learned group whose orientation can be understood by the robot. We can see that there is a suggestive difference between the different group of participants. Although the median performance is almost same in the none vs. the learned group, the range of the mistakes differ. From the data, we can see that more than a quarter

of the participants in the no-reinforcement group made more mistakes than almost anyone in the learned group. Besides that, in spite of having a larger population, the overall range of mistakes of the learned group is smaller than any other group of participants. The group receiving no reinforcement has the largest magnitude of mistakes. To measure the standardized difference we calculated the Cohens D = 1.93 on the 28% of these two populations which signifies that only there is approximately 32% similarilty between both the populations. To show the maximum effectiveness of reinforcers on the learned group only 28% of the data is considered. The reinforcement strategy is considered to be working for a participant when the participant starts making fewer mistakes with same kind of reinforcer, and this also leads to a lower computed regret for the robot.

In the case of the random group, the range of mistakes is smaller than first group because some kind of positive motivation is given out, even if it isn't the most appropriate for the individual. Hence the number of mistakes are also smaller than the first group. From Figure 13 we can tentatively conclude that the people in the learned model condition performed better than the people in the other groups.

Figure 14 shows the fraction of people in all the three populations who made more than three mistakes. Since subjects who performed close to perfectly received little feedback regardless of their experimental category, we would not expect to see much of an effect among those subjects. In this figure, we restrict our attention to subjects who received significant feedback. In this case, a z-test performed between the no-reinforcement and the learned group shows a p-value of 0.03. Thus subjects who made mistakes in the learned group received helpful feedback and improved their performance significantly more than the others.

Reinforcer evaluation

To measure the effectiveness of reinforcers, we calculated the interactions elapse before Baxter realizes which reinforcer is working for a participant. In a few cases it discovered the best reinforcer in the first interaction, but this is not usually the case.

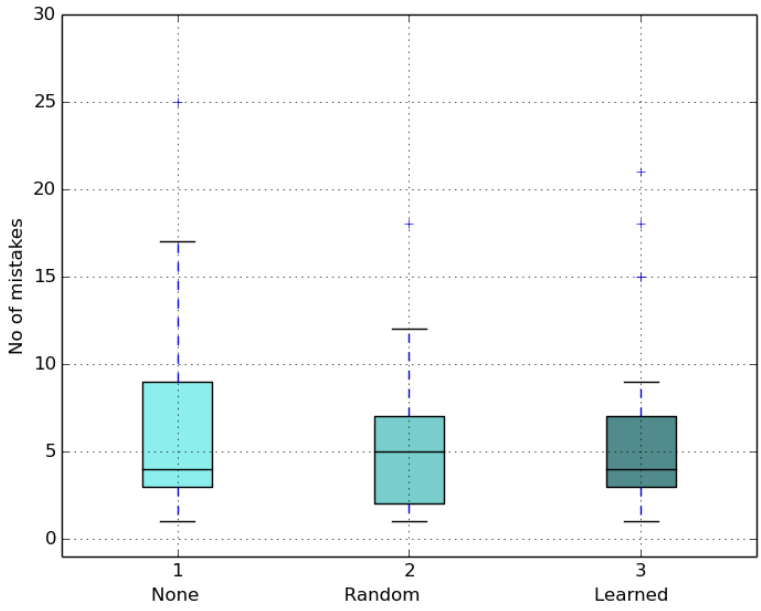

Figure 13. Performance of participants in experiment positive reinforcers with failure.

starts with a random interaction, as they all have equal prior weights. But fairly quickly, one interaction tends to stand out among the others as the strategy which works best for Baxter initially interacts with the participant. We considered cases where Baxter realized in the very first interaction which reinforcer will be working for the participant. Here we also considered the case where subjects received only one reinforcement strategy, which induced them to perform very well through out the rest of the task without making any further mistakes. There are also cases where Baxter identified multiple reinforces which work equally well for a participant after several

interactions. The mean and the standard deviation for the various reinforcers are as follows: Gesture: $\mu = 3.0$, $\sigma = 0.76$, None: $\mu = 2.4$, $\sigma = 1.5$, Reward: $\mu = 3.67$, $\sigma = 3.01$, and Verbal: $\mu = 2.75$, $\sigma = 3.19$. We see that the least effective reinforcement strategy, or at least the one that took the longest to learn for the largest number of participants, was gesture-based. The experiment was performed on college-age subjects. Gestures are usually popular among young children; here it is assumed that the subjects lack emotional engagement, so gestures had less effect than other reinforcement strategies. In case of verbal reinforcement, more interactions were required by the robot to understand the orientation of human participants. This is because Baxter narrates the reinforcer in a machine voice, which is sometimes difficult to comprehend. Subjects encountered some difficulty in understanding and obtaining motivation from verbal interaction. Participants had similar responses to reward and verbal motivations.

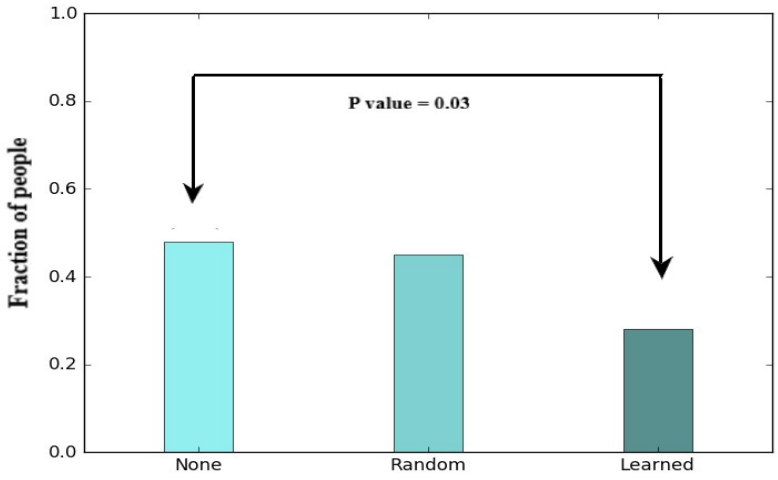

Figure 14. Fraction of participants making more than three mistakes.

Regret analysis

As mentioned in Chapter 3, regret is calibrated on the basis of the decision making ability of the robot. It depends upon the subject's performance, which helps in characterizing the most appropriate reinforcement learning strategy. We correlated the number of mistakes made by the human participants (Figure 15) and the total regret felt by the robot and found a linear relationship between the number of mistakes and the robot's regret. The value of the coefficient is r = 0.88; thus the robot's regret is strongly correlated and the reinforcement learning strategy used by the robot to understand human responses and improve their performance is appropriately working. For the participants who had several interactions with the robot or made many mistakes, Baxter tried to explore different reinforcement strategies at different times, trying to increase their learning rate. Hence we can derive that Baxter can

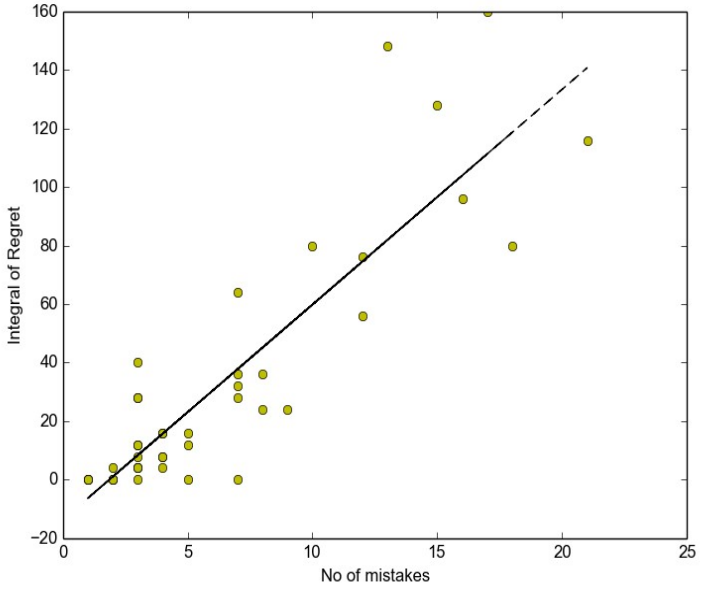

Figure 15. Regret analysis of the robot.

successfully train people to achieve complexes task using their preferred motivations.

Experimental Procedure III

n = 34 (age : $\mu = 19.69$, $\sigma = 3.47$, male=13, female=21, none=11, random=11, learned=12) participants were recruited for the experiments with Baxter, which ran for a time t ($\mu = 18.47, \sigma = 5.60$ in minutes). Among the subjects, 75% had never interacted with a robot, 8.30% interacted a year ago, and 5.50% each a month and a week ago. The task involved in the experiment was divided into two large sections that was further divided into two smaller subsections. The tasks are designed to observe successful skill transfer from robot to human using Simpson's psychometric model, where each category in the taxonomy transitions to another with the goal of skill transfer. In this experiment we only used Guided Response, Mechanism and Adaptation. Guided Response I and II occur in the first half of the experiment where the robot first teaches the participant about the augmented markers and then motivates them throughout the learning process. During this the robot also evaluates the performance of the participants. In the second half of the first section the robot teaches the participant a complex pattern with dual-faced augmented markers and asks the student to reconstruct it. Again, during this process, the robot positively reinforces the learner with simple yes-or-no, random or learned MRL feedback depending upon their assigned experimental group. Participants are allowed to observe the pattern making process and then the markers are immediately shuffled and they are asked to start the reconstruction immediately. Baxter transitions its left hand camera from one spot to another for evaluation. Baxter does not progress to the next position until the participant rectifies any mistakes. During this process the participants get hint, simple, verbal or gesture feedback depending upon their group and Baxter with their performance tries to identify their cognitive orientation. Figure 7 and Figure8 (top frames) corresponds to the experimental procedure with Baxter.

In the second half of the experiment the participants are asked twice to reconstruct the

pattern, this time without any motivation, to observe how well the skill transfer succeeded. In the last half of the experiment participants are asked again to identify two random markers from the set they were taught at the beginning of the experiment to analyze adaptation. Each participant in the experiment was assigned different complex patterns for reconstruction. The marker placement at each position depends upon the robot. At the end of the experiment, the subjects were given questionnaires to answer using a 5-point Likert scale. The results section is further divided into subsections discussing the performance of the participants, the mutual information shared, the robot's regret and the mental model of the participants during the task.

Subjective performance evaluation

To quantify the skill transfer procedure, we calculated the number of mistakes made by each participant in all of the groups in different phases of the experiment. Figure 16 and Figure 17 shows the number of mistakes according to Simpson's psychometric model. There are participants (≈ 8.82%) who didn't commit any mistake throughout the experiment, so their results are not included in the mistakes data. From Table , Figure 16 and Figure 17 we can see that the number of mistakes made during the Mechanism and Adaptation phases are significantly less than during the Guided Response phases irrespective of the groups, which shows the effect of robot feedback during the task. Again if we compare the skill transfer among the groups in the figure, we can see that the number of mistakes in the learned MRL group is comparatively.

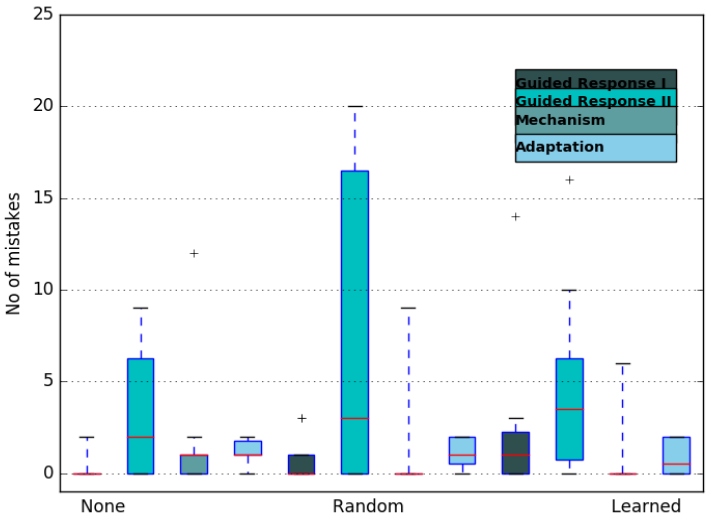

Figure 16. Skill transfer is analyzed different levels of Simpson's psychometric model.

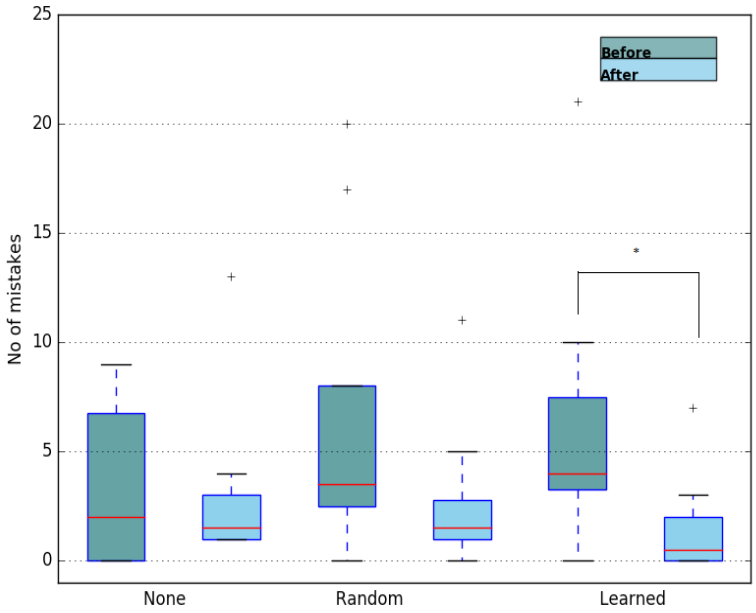

Figure 17. Mistakes by participants before and after the skill transfer with Baxter.

MRL improves skill transfer ($p < 0.05$) lower than the other two. Table 1 presents the performance of the participants in all the phases of the experiment. Figure 17 also plots the

number of mistakes in each group before (Guided Response I and II) and after (Mechanism and Adaptation) the skill transfer procedure; we see that the participants made comparatively fewer mistakes in the learned group than in the other two. Using a t-test for skill transfer outcomes while comparing the learned group with random reinforcers we get a significant p value < 0.05.

Table 1. Performance of the participants in different phases

Group	**None**		**Random**		**Learned**	
	M	SD	M	SD	M	SD
Guided Response I	0.4	0.84	0.8	1.17	2.08	3.92
Guided Response II	2.08	3.92	7.45	8.37	4.5	4.80
Mechanism	**1.8**	3.65	**1.33**	3.04	**0.64**	1.80
Adaptation	**1.2**	0.63	**1.18**	0.87	**0.83**	0.94

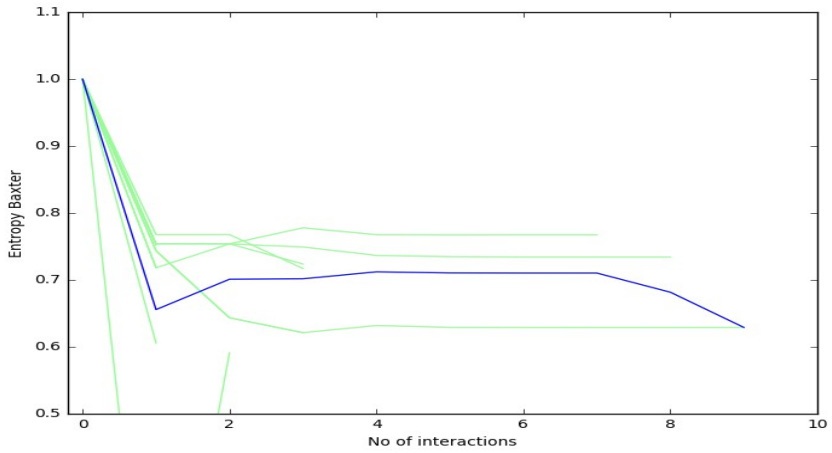

Figure 18. Entropy (green) of the information of robot interacting with participants.

Thus we can conclude that MRL has successfully worked in terms of the skill transfer procedure viz. Guided Response I , Guided Response II, Mechanism and Adaptation using Simpson's psychometric model in the skill transfer scenario in Baxter.

Entropy analysis

Entropy denotes the randomness of a system. In Figure 18, we can see that the entropy of the information of the robot obtained by interactions goes down monotonically along with each interaction with its human participant in MRL. With each interaction, the robot is gaining more information about the participants' performance and their cognitive orientation towards each reinforcer. The pale green lines show the entropy of the information of the each participant obtained by interactions with Baxter, and the blue line is their mean performance. Not every participant had an equal number of interactions with Baxter, but regardless, using Algorithm 1, the machine manages to gain information steadily about their performance. For each participant, the value of entropy varied depending on the exact pattern of the robot's choices of exploration and exploitation, but we can conclude that for each participant it has gained some information at each interaction which later helped it to construct a successful mental model.

Regret analysis

As mentioned in Chapter 3, regret is calibrated on the basis of the decision making ability of the robot. It depends upon the subject's performance, which helps in characterizing the most appropriate reinforcement learning strategy. We correlated the number of mistakes made by the participants and the total regret felt by the robot (Figure 19) and found a linear curve

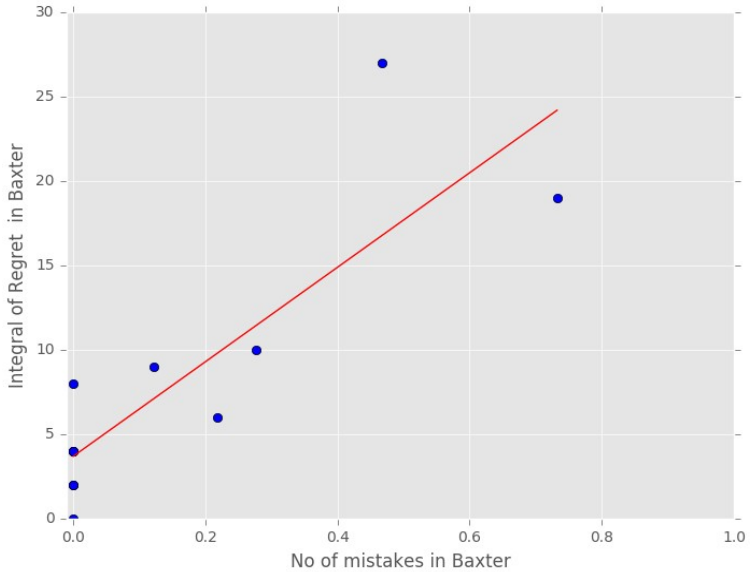

Figure 19. Regret analysis of Baxter.

shows the best fit with coefficient r = 0.85. linear relationship between the number of mistakes and the robot's regret. The value of the coefficient is r = 0.85; thus the robot's regret is strongly correlated and the reinforcement learning strategy used by the robot to understand human responses and improve their performance, is working appropriately. For the participants who had several interactions with the robot or made many mistakes, Baxter tried to explore different reinforcement strategies at different times, trying to increase their learning rate. This illustrates that Baxter can successfully train people to achieve complex task using their preferred motivations.

Mental model analysis

At the end of the experiment, those participants in the learned model group were asked to choose their preferred reinforcers. Baxter could correctly identify the preferred reinforcers in half of the

cases (twice as effectively as a random baseline). Thus, MRL allowed the robot successfully to identify the cognitive orientation of the participants to a large extent (accuracy score = 0.50). During the task, since the number of interactions was limited, the robot did not have sufficient opportunity to engage in the exploitation aspect of the reinforcement learning, and thus its ability to identify preferred reinforcers was limited (but still reasonably successful). In these experiments, Baxter explored more than exploited, which impacted the types of reinforcers given out by the robot.

At the end of each experiment we probed participants with a 5-point Likert scale(1: Strongly disagree; 2: Disagree; 3: Neutral; 4: Agree; 5: Strongly agree). Participants 39in the no reinforcer group (μ=3.54 , σ=0.68), random reinforcer group (μ=4 , σ=0.77), learned reinforcer group (μ=3.25 , σ=1.22) wanted to play with Baxter again and again in no reinforcer group (μ=4.0 , σ=0.63), random reinforcer group (μ=3.63 , σ=0.92) and learned reinforcer group (μ=2.6 , σ=0.89) thought it is useful as a teacher. Interestingly, the people who performed poorly during the experiment neither found Baxter to be useful nor thought it was a good teacher. In other words, people who enjoyed the interaction also found it helpful and wanted to come again to learn from the robot, whereas those were not fond of the robot were not interested in the experiment and ended up performing poorly.

Experimental Procedure IV

n = 31 (age μ = 19.77, σ = 5.07, male=9, female=22, none=11, random=10, learned=10) participants were recruited for the experiments with Tetris for 15 minutes. We conducted an experiment in a gaming scenario to observe the performance of MRL across different platforms. Among all the participants, 74.19% of the subjects had played the game but not within the last

year, 12.19% had played within the last year but not the last month, and the rest had played the game within the last month. Like with the Baxter scenario, participants were also being trained to be better Tetris players. The skill transfer scenario is also analyzed with Simpson's psychometric model (Guided Response and Adaptation).

As is common in Tetris games, during each move, the next block is shown alongside the 10x20 game board so that players plan their move ahead. As in the previous experiment, the teaching process is divided into two phases. Initially the participants are asked to play the game for 15 mins with reinforcers provided depending upon their assigned experimental group (simple, random, MRL).

Whenever the participant makes a mistake, the machine alerts them, provides are reinforcer, and then they are allowed to continue. For Tetris, mistakes are considered to be placing a block in such a way to hinder fast scoring. Wrong placement is associated with forming a gap between lines which will make it difficult to eliminate the line of blocks in the future. Also, if a player places blocks several times in a row without eliminating any lines, that is also considered as a mistake as the towers of blocks build up and get closer to ending the game. During the first experiment task, if the game is lost, participants are allowed to restart, so that all subjects received the same time to learn properly. In the second portion Figure 20. Scores acquired by participants per minute before and after skill transfer during Tetris. In the random group, subjects showed significant progress after skill transfer $p<0.05$. of the experiment (Adaptation), participants are asked again to play the game, this time without reinforcement training, to assess the quality of their learning. They are asked to play until losing, or until 5 minutes elapsed ($\mu=2.68$ $\sigma=0.88$ minutes). Since the experiment had only two phases, we used Guided Response and Adaptation to analyze the skill transfer. After the game, the subjects were also given questionnaires probed with a 5-point Likert scale. Similarly to the Baxter experiment, our

analysis of the Tetris scenario is divided into subjective and entropy evaluation, regret and mental model analysis.

Subjective performance evaluation

For Tetris, we computed the skill transfer on the basis of Simpson's psychometric model. Since, in a game like Tetris, people are expected to make a large number of mistakes, we computed the scores per minute of the participants of different groups. Figure 20 denotes the scores of the participants before (Guided Response) and after (Adaptation) the skill transfer procedure. The subjects in the random group score significantly better than the

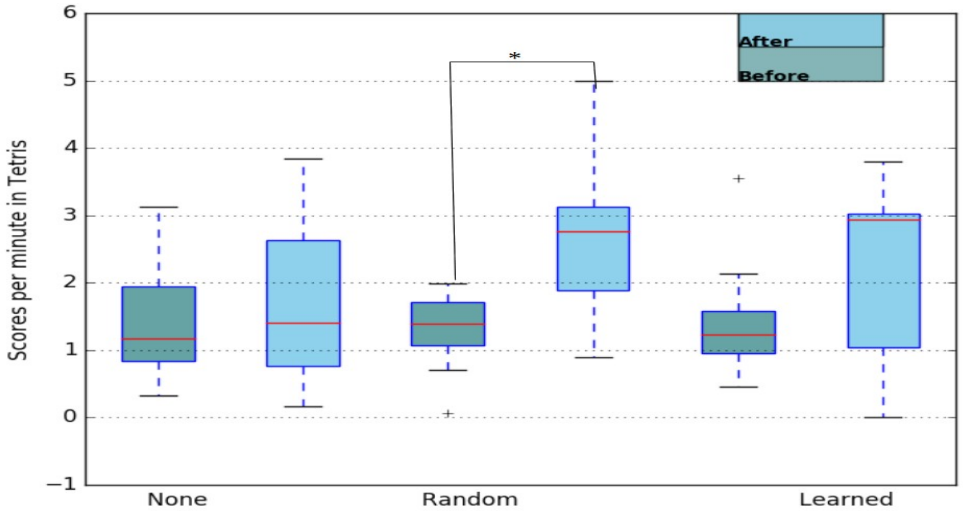

Figure 20. Scores by participants per minute before and after skill transfer during Tetris.

Here in case of the random models the participants got some sort of feedback which reinforced them to perform better at the task. The random group showed significant improvement than the none group ($p \approx 0.01$). Hence although the participants

responded to feedbacks , there is not much difference between the random and the learned models. Since in this task may people have already played Tetris before the MRL didn't have salutary effect on the participants.

Entropy analysis

Like Baxter, this shows that the computer is gaining information about the participant's performance and its cognitive orientation towards each reinforcer with each interaction.

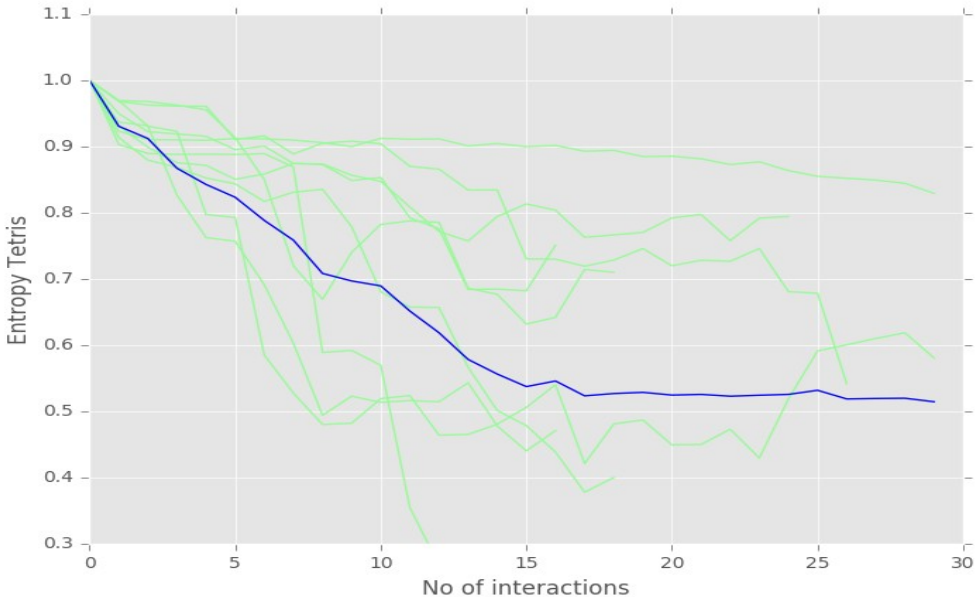

Figure 21. Entropy (green) of information of Tetris while interacting with participants.

The entropy of the gaming device's information monotonically decreases with each interaction with its human participant in MRL. Like Baxter, this shows that the computer is gaining information about the participant's performance and its cognitive orientation towards

each reinforcer with each interaction. The pale green lines show the entropy of each participant over each interaction with the machine and the blue line shows the mean performance.

Regret analysis

Like Baxter, here also we tried to calibrate the regret of the gaming system. The correlation coefficient in this case is $r < 0.50$. Although the system tried to explore different reinforcement strategies at different times, trying to increase the learning rate, the results are suggestive and not conclusive. Figure. 22 is the best fit curve to analyze the relation between integral of regret and number of mistakes in Tetris.

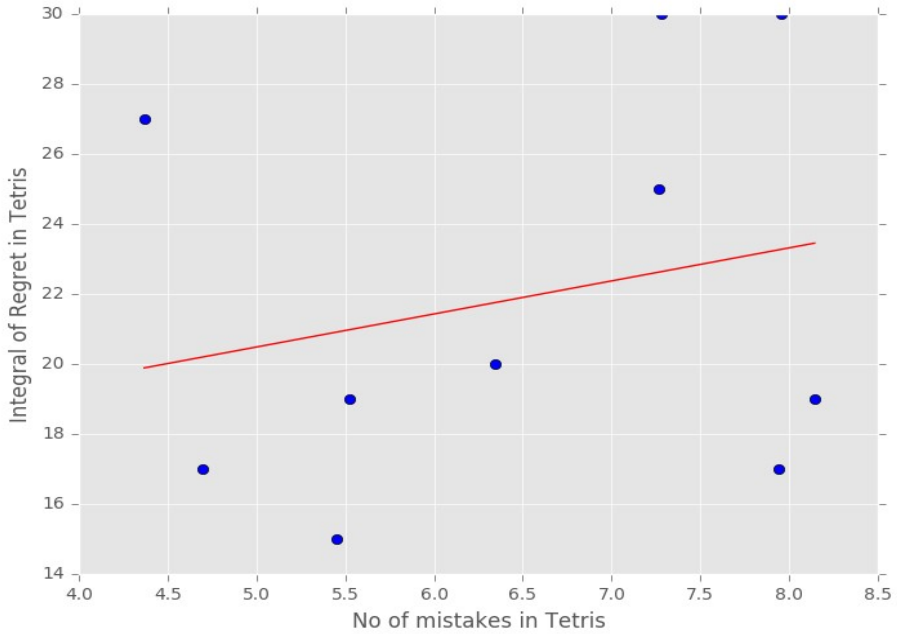

Figure 22. Regret analysis of Tetris.

The linear curve shows the best fit with coefficient $r < 0.50$ regret of the system and the integral of the number of mistakes made by the participants. This is because the participants utilized all the reinforcers given to them without distinguishing between them very much.

Mental model analysis

At the end of the experiment, those participants in the learned model were asked to choose their preferred reinforcers. The machine could partially identify the preferred reinforcers in Tetris. For example, reinforcer a is often misclassified as e in many cases. In Tetris, since all the reinforcers were fairly similar to one another, in spite of the fact that they guided participants differently, the subjects often failed to distinguish the efficacy of one versus another(accuracy score < 0.50). Hence they were able to make use of all of the provided reinforcers successfully, without much differentiation between them. It was difficult for the system to determine the preferred reinforcer of the participant, and only it occasionally successfully identified the best reinforcers. Hence there is not much difference between the performance of the learned group with that of the random group of participants. At the end of each experiment we probed participants with a 5-point Likert scale (1:Not all helpful; 2: Moderately helpful; 3: Neutral; 4: Helpful; 5: Extremely helpful). The subjects wanted to play Tetris again with (μ=3.0, σ= 0.70) none, (μ=3.2, σ= 0.78) with random and (μ=3.3, σ= 0.82) with MRL. They thought it was useful (μ=2.1, σ= 0.75) with none, (μ=2.5, σ= 0.85) with random and (μ= 2.5, σ=0.97) with MRL. Overall, more than half the subjects found the reinforcers useful and therefore thought the system was a good teacher. These statistics show significant results.

CHAPTER VI

DISCUSSION

The subjects provided with semantic structure happened to have lower mean joystick proficiency than the other group. We hypothesized that subjects learning with semantic labels would retain the task better than the other group, although we did not quite confirm this to a statistically significant level, though we can still say that the robot made a decent teacher. Since all of the subjects in both groups were allowed to take copious notes and use them during the experiment, we believe that the task execution challenge became too easy, accounting for the relative success of the non-semantic structure group. In addition, the chosen sign language phrase may not have been sufficiently complex to differentiate the learning process. Even so, the subjects in the no semantic structure group did indeed perform less well than those with access to human-accessible semantic guidance. Subjectively, participants without semantic labels more often skipped trajectories and chose incorrect motions more often. Since they were novice operators, they were not precise about their movements. For example, in the first two movements, the robot's arm is aligned perpendicularly with its shoulder in the ideal motion. However, the participants sometimes failed to achieve this pose, which caused significant deviation since the shoulder joint influences the subsequent position of all other joints. The subjects had little understanding of how precise various actions needed to be, and this was true for both groups.

During the experiment, displayed on a monitor placed beside each candidate. It is likely the case that the learning and teaching would have gone better with a more audiovisual interaction. Without any audio, it was difficult for some participants to keep track of the labels on the screen and Baxter's movement at the same time. Especially during fine trajectory adjustments, it was very difficult to look at both the robot and the monitor simultaneously. These are a few of the factors which might have influenced the performance of the participants during the experiment irrespective of their group.

In the experiment, the results are sometimes not as strong as we might hope for several reasons. If the robot's grippers were closed, they occasionally hindered the camera, blocking the robot from identifying the markers, since we used the left hand camera for detection and evaluation purposes. The Baxter arm and gripper are not extremely dextrous; it is sometimes very unsophisticated in its attempts to pick up the blocks whenever they not lying perpendicularly to the camera. The markers, after several tasks and the degradation which resulted from repeated handling by both human and robot, became unclear and difficult for the marker tracking algorithm to recognize, which also contributed to system crashes. Again, many of the young adults who participated in the experiment failed to connect to the robot emotionally and lacked engagement. Some subjects paid very little attention to the robot's attempts to communicate a reinforcement strategy, to the point that the subjects attempted to interact with the researchers conducting the experiment rather than the robot. Some participants simply produced iteration after iteration of patterns until they happened upon the correct one, without paying attention to the robot gamely attempting to help. Rather, they simply tried each block at each position to figure out the right approach. Hence Baxter on its end was confused in providing the reinforcement strategy. For this reason, we see that Baxter was only able to identify a successful motivational strategy for half the participants in the learned group.

Another potential point of alienation came from the fact that Baxter's voice did not issue from the robot itself, but rather a speaker off to the side (since the robot hardware lacks sound capability). The students had to turn to their right side and interact with a computer console to give Baxter their feedback in form of yes or no, which is unsophisticated; verbal interaction would have been a better option. However, participants in the experiment came from different national backgrounds and language abilities, so it was very hard for the robot to understand their pronunciation, and we were therefore forced to keep the human feedback in that format.

Seven blocks with 14 markers can be placed in many, many ways, but no participants required nearly that many attempts to figure out the correct pattern. Thus, even when they did not directly engage with the robot's attempts to teach, it still had some impact on them. We used the spatial arrangement of markers as our complex task. Some people who performed better in all the groups might simply be good at this style of task and would fail at some other complex task. The difference in performance between the groups might be different in different complex task scenarios. Our MRL theory applies to the people who have performed poorly in the task, and therefore received appropriate motivations. They might be better at a different complex task, and therefore engage differently with the reinforcement behavior. In the case of Tetris, people are well acquainted with the game, so the reinforcer might only have a little effect on them. Although the autonomous agents successfully developed a teaching strategy for only half of the participants, it is enough to suggest that such feedback does have impact on human behavior and learning. Furthermore, this approach allows the autonomous agent to assess its own success and learn to calibrate its own interactions in ways that lead to successful teaching.

CHAPTER VII

CONCLUSION

In this work, we studied the problem of skill transfer from a robot to human, where the autonomous agent is not only learning about the human mental model but also trying to adapt its own accordingly. In this shared environment, the robot is trying to maximize the cumulative reward by learning about human behavior and simultaneously improving its own cognitive model. We highlighted mutual information communicated between the robot and the human, and validated their interaction in skill transfer using real-time experiments in both robot and gaming platforms. The subjective performance, information gain over time and the confusion matrices give us a conclusive idea how robots and computer systems can successfully transfer skills from themselves to humans. In our future work we would like to implement MRL across different platforms. Heavy construction equipment like excavators and backhoes are required to perform complex tasks like digging, truck loading and ditch crossing, requiring a series of complex manipulations. Learning appropriate manipulations for these different situations is a hard task. We want to implement MRL in these scenarios where humans can learn the subtlety of control manipulations with robot assistance. In addition, we intend to investigate the necessary behavioral changes required to be adapted by the robots to become better trainers over time. We would also like to involve robots in guiding students towards correct actions and simultaneously identifying their mental models in a robot-human interaction.

BIBLIOGRAPHY

Abbeel, P., & Ng, A. Y. (2004, July). Apprenticeship learning via inverse reinforcement learning. *In Proceedings of the twenty-first international conference on Machine learning* (p.1).

Audibert, J. Y., Munos, R., & Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19), 1876-1902.

Audibert, Jean-Yves, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* 410.19 (2009): 1876-1902.

Baranes, A., & Oudeyer, P. Y. (2009). R-iac: Robust intrinsically motivated exploration and a ctive learning. *IEEE Transactions on Autonomous Mental Development*, 1(3), 155-169.

Bickhard, M. H. (2009). Interactivism: A manifesto. *New Ideas in Psychology*, 27(1), 85-95.

Boccanfuso, L., Barney, E., Foster, C., Ahn, Y. A., Chawarska, K., Scassellati, B., & Shic, F. (2016, March). Emotional robot to examine different play patterns and affective responses of children with and without ASD. In 2016 *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 19-26). IEEE.

Cakmak, M., & Thomaz, A. L. (2012, March). Designing robot learners that ask good questions. *7th ACM/IEEE International Conference on Human-Robot Interaction* (HRI) (pp. 17-24). IEEE.

Cakmak, M., Chao, C., & Thomaz, A. L. (2010). *Designing interactions for robot active learners. IEEE Transactions on Autonomous Mental Development*, 2(2), 108-118.

Cakmak, M., DePalma, N., Thomaz, A. L., & Arriaga, R. (2009). Effects of social exploration mechanisms on robot learning. In RO-MAN 2009-*The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 128-134). IEEE.

Chan, L., Hadfield-Menell, D., Srinivasa, S., & Dragan, A. (2019, March). The assistive multi-armed bandit. *In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 354-363). IEEE.

Cheung, E. O., & Gardner, W. L. (2015). The way I make you feel: Social exclusion enhances the ability to manage others' emotions. *Journal of Experimental Social Psychology*, 60, 59-75.

Clement, B., Roy, D., Oudeyer, P. Y., & Lopes, M. (2013). Multi-armed bandits for intelligent tutoring systems. *arXiv preprint arXiv:1310.3174*.

Crick, C., Osentoski, S., Jay, G., & Jenkins, O. C. (2011, March). Human and robot perception in large-scale learning from demonstration. *In Proceedings of the 6th international conference on Human-robot interaction* (pp. 339-346).

Dominey, P. F., & Warneken, F. (2011). The basis of shared intentions in human and robot cognition. *New Ideas in Psychology,* 29(3), 260-274.

E Simpson. *The classification objectives in the psychomotor domain*–washington. Gryphon House, 1972.

Fan, Y., Tian, F., Qin, T., Li, X. Y., & Liu, T. Y. (2018). *Learning to teach*. arXiv preprint arXiv:1805.03643.

Fasola, J., & Matarić, M. J. (2013). *A socially assistive robot exercise coach for the elderly. Journal of Human-Robot Interaction*, 2(2), 3-32.

Ferreira, E., & Lefevre, F. (2015). Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards. *Computer Speech & Language*, 34(1), 256-274.

Fitter, N. T., & Kuchenbecker, K. J. (2016, November). Designing and assessing expressive open-source faces for the Baxter robot. *In International Conference on Social Robotics* (pp. 340-350). Springer, Cham.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., & Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. *In Advances in neural information processing systems* (pp. 2625-2633).

Grizou, J., Lopes, M., & Oudeyer, P. Y. (2013, August). Robot learning simultaneously a task and how to interpret human instructions. *In 2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)* (pp. 1-8). IEEE.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *In Advances in neural information processing systems* (pp. 3909-3917).

Kaelbling, L. P., & Lozano-Pérez, T. (2011). Hierarchical task and motion planning in the now. *In 2011 IEEE ICRA*, 1470–1477.

Keskin, S. C. (2014). From what isn't empathy to empathic learning process. *Procedia-Social and Behavioral Sciences*, 116, 4932-4938.

Khan, F., Mutlu, B., & Zhu, J. (2011). How do humans teach: On curriculum learning and

teaching dimension. *In Advances in neural information processing systems* (pp. 1449-1457).

Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., & Scassellati, B. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *Journal of autism and developmental disorders*, 43(5), 1038-1049.

Knox, W. B., & Stone, P. (2008, August). Tamer: Training an agent manually via evaluative reinforcement. *In 2008 7th IEEE International Conference on Development and Learning* (pp. 292-297). IEEE.

Knox, W. B., & Stone, P. (2009, September). Interactively shaping agents via human reinforcement: The TAMER framework. *In Proceedings of the fifth international conference on Knowledge capture* (pp. 9-16).

Knox, W. B., Stone, P., & Breazeal, C. (2013, October). Training a robot via human feedback: A case study. *In International Conference on Social Robotics* (pp. 460-470). Springer, Cham.

Koenig, N., Takayama, L., & Matarić, M. (2010). Communication and knowledge sharing in human–robot interaction and learning from demonstration. *Neural Networks*, 23(8-9), 1104-1112.

Konidaris, G., Kuindersma, S., Grupen, R., & Barto, A. (2012). Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, 31(3), 360-375.

Konok, V., Korcsok, B., Miklósi, Á., & Gácsi, M. (2018). Should we love robots?–The most liked qualities of companion dogs and how they can be implemented in social robots. *Computers in Human Behavior*, 80, 132-142.

Lee, S. L., Lau, I. Y. M., Kiesler, S., & Chiu, C. Y. (2005, April). Human mental models of humanoid robots. *In Proceedings of the 2005 IEEE international conference on robotics and automation* (pp. 2767-2772). IEEE.

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2012, October). Long-term interactions with empathic robots: Evaluating perceived support in children. *In International Conference on Social Robotics* (pp. 298-307). Springer, Berlin, Heidelberg.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. *International journal of human-computer studies*, 71(3), 250-260.

Litoiu, A., & Scassellati, B. (2015, March). Robotic coaching of complex physical skills. *In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (pp. 211-212).

Lockwood, P. L. (2016). The anatomy of empathy: Vicarious experience and disorders of social cognition. *Behavioural brain research*, 311, 255-266.

Lopes, M., Melo, F., & Montesano, L. (2009, September). Active learning for reward estimation in inverse reinforcement learning. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 31-46). Springer, Berlin, Heidelberg.

Mihaela, P. (2013). The structure and dynamics of the teacher's empathic behavior. *Procedia-Social and behavioral sciences*, 78, 511-515.

Nikolaidis, S., & Shah, J. (2013, March). Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. *In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 33-40). IEEE.

Nunez, E., Matsuda, S., Hirokawa, M., & Suzuki, K. (2015, October). Humanoid robot assisted training for facial expressions recognition based on affective feedback. *In International Conference on Social Robotics* (pp. 492-501). Springer, Cham.

Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G., & Breazeal, C. (2017, March). Growing growth mindset with a social robot peer. *In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 137-145).

Powell, P. A., & Roberts, J. (2017). Situational determinants of cognitive, affective, and compassionate empathy in naturalistic digital interactions. *Computers in Human Behavior*, 68, 137-148.

Prozesky, D. R. (2000). Teaching and learning. *Community eye health*, 13(36), 60.

Ramachandran, A., & Scassellati, B. (2015, September). Developing adaptive social robot tutors for children. *In 2015 AAAI Fall Symposium Series*.

Roy, S., Crick, C., Kieson, E., & Abramson, C. (2018, August). A reinforcement learning model for robots as teachers. *In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 294-299). IEEE.

Roy, S., Kieson, E., Abramson, C., & Crick, C. (2018, March). Using human reinforcement learning models to improve robots as teachers. *In Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 225-226).

Roy, S., Kieson, E., Abramson, C., & Crick, C. (2019, March). Mutual reinforcement learning with robot trainers. *In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 572-573). IEEE.

Roy, S., Maske, H., Chowdhary, G., & Crick, C. (2017, March). Teaching and learning using semantic labels. *In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 271-272).

Salter, T., Dautenhahn, K., & te Boekhorst, R. (2006). Learning about natural human–robot interaction styles. *Robotics and Autonomous Systems*, 54(2), 127-134.

Sauppé, A., & Mutlu, B. (2015). Effective task training strategies for human and robot instructors. *Autonomous Robots*, 39(3), 313-329.

Scassellati, B. M. (2001). Foundations for a Theory of Mind for a Humanoid Robot (*Doctoral dissertation, Massachusetts Institute of Technology*).

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. *Annual review of biomedical engineering*, 14, 275-294.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3), 379-423.

Simpson, E. J. (1966). *The classification of educational objectives,* psychomotor domain.

Spaulding, S., Chen, H., Ali, S., Kulinski, M., & Breazeal, C. (2018, July). A social robot system for modeling children's word pronunciation: Socially interactive agents track. *In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 1658-1666). International Foundation for Autonomous Agents and Multiagent Systems.

Stojanov, G., Trajkovski, G., & Kulakov, A. (2006). Interactivism in artificial intelligence (AI) and intelligent robotics. *New Ideas in Psychology*, 24(2), 163-185.

Strohkorb, Sarah, and Brian Scassellati. "Promoting collaboration with social robots. *In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: *An introduction. MIT press*.

Tabrez, A., Agrawal, S., & Hayes, B. (2019, March). Explanation-based reward coaching to improve human performance via reinforcement learning. *In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 249-257). IEEE.

Wong, M. M., & Csikszentmihalyi, M. (1991). Motivation and academic achievement: The effects of personality traits and the duality of experience. *Journal of Personality*, 59(3), 539-574.

Wu, C., Lenz, I., & Saxena, A. (2014, July). Hierarchical Semantic Labeling for Task-Relevant RGB-D Perception. *In Robotics: Science and systems*.

Yi, X., Sun, M., Li, R., & Li, W. (2018). Automatic poetry generation with mutual reinforcement learning. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3143-3153).

Yin, H., Billard, A., & Paiva, A. (2015, March). Bidirectional learning of handwriting skill in human-robot interaction. *In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (pp. 243-244).

VITA

Sayanti Roy

Candidate for the Degree of

Doctor of Philosophy

Dissertation: MUTUAL REINFORCEMENT LEARNING TO IMPROVE ROBOTS AS TRAINERS

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for Doctor of Philosophy in Computer Science at Oklahoma State University, Stillwater Oklahoma in May, 2020.

Completed the requirements for Bachelor of Technology in Electronics and Communication Engineering at West Bengal University of Technology, Kolkata West Bengal in 2014.

Experience:

Employed by Oklahoma State University in the position of Research and Teaching Assistant in Stillwater, Oklahoma from August 2015 to date.