Learning Polite Behavior with Situation Models

Rémi Barraquand
INP Grenoble
INRIA Grenoble Research Center
655 Ave de l'Europe, 38334 St. Ismier, France
Remi.Barraquand@inrialpes.fr

James L. Crowley
INP Grenoble
INRIA Grenoble Research Center
655 Ave de l'Europe, 38334 St. Ismier, France
James.Crowley@inrialpes.fr

ABSTRACT
In this paper, we describe experiments with methods for learning the appropriateness of behaviors based on a model of the current social situation. We first review different approaches for social robotics, and present a new approach based on situation modeling. We then review algorithms for social learning and propose three modifications to the classical Q-Learning algorithm. We describe five experiments with progressively complex algorithms for learning the appropriateness of behaviors. The first three experiments illustrate how social factors can be used to improve learning by controlling learning rate. In the fourth experiment we demonstrate that proper credit assignment improves the effectiveness of reinforcement learning for social interaction. In our fifth experiment we show that analogy can be used to accelerate learning rates in contexts composed of many situations.

Categories and Subject Descriptors
I.3.6 [**Learning**]: Robot Learning, Behavioral learning, Learning for Man Machine Interaction, Social Robotics.

General Terms: Algorithms, Experimentation

Keywords: Social Interaction, Social Learning, Social Robotics, Q-Learning, Credit assignment, Learning by Analogy.

1. Situated Social Common Sense
With current technology, systems and services are unable to discriminate between appropriate and inappropriate behaviors. As a result, most attempts at proactive user services produce systems that are highly disruptive of human activity. In short, computing systems lack social common sense.

Common sense is the collection of shared concepts and ideas that are accepted as correct by a community of people. Social common sense refers to the shared rules for polite, social interaction that implicitly rule behavior within a social group. To a large extent, such common sense is developed using implicit feedback during interaction between individuals. Our goal in this research is to develop methods to endow an artificial agent with the ability to acquire social common sense using the implicit feedback obtained from interaction with people. We believe that such methods can provide a foundation for socially polite man-machine interaction, and ultimately for other forms of cognitive abilities.

HRI'08, March 13-15, 2008, Amsterdam, Netherlands.

In this paper, we propose to focus on a key aspect of social common sense: the ability to act appropriately in social situations. In this work, we have sought to train an association between behavior and social situation. Our approach for modeling social situations is inspired by the cognitive models for situation proposed by Johnson-Laird [14] in which situations are modeled as relations between entities. In previous work, we have generalized situation models with the introduction of the concept of "role" [8] and experimented with the use of machine learning techniques for automatically acquiring situation models. In this paper we extend this approach to the problem of learning social common sense. It is our intention that these methods may be used with any system that can spontaneously act to propose information or services.

Rules for polite social interaction tend to be highly dependent on context, as well as specific to individuals or groups. Thus we have sought methods that would allow systems to learn the appropriateness of actions using the natural social feedback that people provide in most social contexts. Although we have used the sensory channels provided by an AIBO robot, these methods may be used with any techniques that enable machine perception of human social signals.

In the following section we review previous approaches to social robotics, and then describe the situation modeling method on which our approach is based. In section 3 we describe algorithms for online learning from social feedback. We adopt an approach based on reinforcement learning, and propose a modified learning algorithm that allows for proper credit assignment. In section 4 we describe our experimental set up and methods. In section 5 we present preliminary experiments that show that proper credit assignment greatly improves the effectiveness of reinforcement learning, and that analogy can be used to greatly accelerate learning rates.

2. Research on Social Interaction
A technology for polite social interaction can draw on concepts and models from a diverse variety of fields. While much work is currently concentrated within robotics, relevant concepts may also be found in the social and cognitive sciences.

2.1 Social Robotics
Over the last decade a growing sub-community within the field of autonomous robotics has turned its attention to the problem of constructing social robots [6, 12, 16]. Fong et al. [11] reviews socially interactive robots, and discuss different forms of social robots as well as potential contributions from other research domains. They present design methods and describe the potential impact of such robots on humans. In particular, they claim that social interaction requires that systems be responsive to the non-linguistic signals that human exhibit in human-to-human

interaction, including eye-gaze, turn-taking, theory of mind and imitation.

Breazeal [6] proposes a hierarchy of four classes of social robots, from socially evocative to sociable. As one moves progressively up the hierarchy, the abilities of the robot to engage in social interaction increase. Within this hierarchy, socially evocative robots are designed to encourage people to anthropomorphize the technology in order to interact with it. Socially communicative robots use human-like social cues and communication modalities to facilitate interactions with people. Socially responsive robots are able to learn from their interaction and social partners. Sociable robots are socially participative, and maintain their own internal goals and motivations.

The robot Kismet [7] is an anthropomorphic robot that engages people in natural and expressive face-to-face interaction. Kismet is inspired by infant social development, psychology, and ethnology, and belongs to the class of sociable robots defined by Breazeal. To imitate human abilities, Kismet has been provided with visual feature extraction, an attention system, a perceptual system, a motivation system, a behavior system and a motor system. Kismet has been designed to support and develop social cues and skills that could ultimately play an important role in socially-situated learning with a human instructor. Kismet provides a powerful tool to study and develop social interaction capabilities, because it uses models for human-to-human interaction to imitate human interaction mechanisms [1].

The Roboceptionist project [12] has addressed the problem of learning from continued long-term interaction over periods of days, weeks, and even years. This project sought to provide useful services depending on the situation, and also to exhibit personality and character. The robot is designed to be compelling enough to encourage multiple visits over extended periods of time, and to encourage interaction with non-experts. The results obtained are claimed to be encouraging, and researchers are now working toward making its behavior more human-like, in an effort to improve the quality of its interactions with visitors.

Recognizing emotional state is often considered necessary for natural interaction. Several authors report methods to recognize emotional states such as despair, happiness and boredom using face expressions [20, 23]. For example, Bartlett [2] proposes a real time algorithm to find a face in an image and to recognize facial expression, implemented on a Sony's AIBO robot. Breazeal [5] explored the use of voice to recognize the communicative intent of a partner in an interaction. De Silva [9] uses both voice and face expression to recognize emotion.

Isbell et al [13] reports on the use of reinforcement learning with a software agent. They describe an agent named Cobot that can proactively execute actions in a multi-user chat environment and adapt its behavior from multiple sources of human reward. After 5 months of training, and 3171 reward and punishment events from 254 different users, Cobot learned nontrivial preferences for a number of users, modifying its behavior based on its current state. This is an early approach of the use of reinforcement learning in a complex human online social environment, where many of the standard assumptions (stationary rewards, Markovian behavior, and appropriateness of average reward) are not valid.

The results obtained with Cobot illustrate that Reinforcement Learning can be used in open-ended social settings. However, the appropriateness of a behavior is highly dependent on the current social situation. Without some means to model the social situation, learned behaviors are likely to be applied in an inappropriate manner.

We propose to capture social common-sense by training the appropriateness of behaviors in social situations. A key challenge is to employ an adequate representation for social situations.

2.2 Situation Models for Social Interaction

Situation models have been proposed by Johnson-Laird [15] as a cognitive theory for human mental-models. While his model, as well as much of the subsequent literature in this area, has been concerned with spatial reasoning or linguistic understanding [14], these concepts can be adopted for the construction of software systems and services for understanding social interaction. In previous work, [8] we have described the use of situation models for context aware services. As in the cognitive modeling literature, situations are defined as a set of relations between entities, where entities may be agents, objects or any abstract concepts observed as a correlated set of properties. In our model, situations are organized into networks, with transition probabilities, so that possible next situations may be predicted from the current situation. This model has been used to construct a variety of services including services for recording events in a meeting or lecture, privacy protection services and other communications services.

A situation model has two facets: perception and action. Brdiczka [3, 4] has provided methods for learning entities, roles, situations and state transitions for recognizing situations. In the work described here, we assume that a situation model has been provided and concentrate on learning the appropriateness of possible actions or behaviors that may be chosen in each situation.

Attention is an important relation in social situations. According to Maisonnasse [18], attention is defined as a process of concentrating cognitive resources on selected aspects of the environment while ignoring others. In our approach, we adopt the attention model developed by Maisonnasse to include the shared attention of agents (human and artificial) as a relation for describing social situations.

3. Reinforcement Learning

Reinforcement Learning methods are commonly used for systems that need to learn from self-generated experience over time. In a standard reinforcement-learning model, an agent is connected to the environment via perception and action channels. At each step *t*, the agent receives some indication of the current state of the environment s_t and chooses an action a_t. This action then changes the environment state and the value of this state transition is communicated to the agent through a scalar reinforcement signal r_t (the reward). The agent seeks to choose actions that tend to increase the long-run sum of values of the reinforcement signal and learn to do so over time by systematic trial and error. This requires finding an optimal policy π as a function π: $s \rightarrow a$, that maps state to action, and a value function that maps the state (or the state and action) to a measure of long-term value.

There are several families of Reinforcement Learning algorithms. Temporal Difference learning allows learning of a value and a policy function by interacting on-line with an environment. The Q-learning algorithm [29] is a form of temporal difference learning in which the value function is defined by

$$Q^*(s_t, a_t) = E\left[R(s_t, a_t) + \gamma \max_{a'} Q^*(s_{t+1}, a')\right]$$

This represents the expected value of the reward for taking action a_t from state s_t, ending up in state s_{t+1}, and then acting optimally from then on. The parameter γ is the weight of the expected reward. The Q-function is typically stored in a table, indexed by state and action. Starting with arbitrary values, the optimal Q-function can be iteratively approximated based on observations.

Interactive Reinforcement Learning (IRL) [6] is an approach for training by natural interaction. Unlike traditional reinforcement learning algorithms, in which the reward signal is only determined based on world state and agent action, with IRL the reward depends on real-time interaction with a human teacher. In an IRL session, the human can choose to change the reward signal not only at certain goal states, but also continuously throughout the interaction.

An important challenge for IRL is to allow humans to remain passive when appropriate, usually deferring to an independent environmental-based reward signal. Thomaz et al. [26, 27] present a framework for studying the role that real-time human interaction plays in training robots to perform new tasks. Socially Guided Machine Learning (SG-ML) assumes that people will train machines through a collaborative process and will expect machines to engage in social forms of learning. They demonstrate that guidance, as well as an asymmetric interpretation of feedback, can accelerate convergence with respect to traditional Q-Lambda. In this case, the reward is not only given to encourage or punish an action, but also to guide the learner in the learning process and thus reduce the exploration time.

Q-Learning provides a standard and widely understood formulation of Reinforcement Learning. In the following, we compare five variations of Q-Learning for use with Social Learning.

3.1 Q-Learning

The Q-Lambda algorithm [29] is a model free method that estimates the state-action value function as follow:

$$Q_{t+1}(s_t,a_t) = (1-\alpha_t)\cdot Q_t(s_t,a_t) + \alpha_t\left(R(s_t,a_t) + \gamma \max_{a'} Q_t(s_{t+1},a')\right)$$

where s_{t+1} is the state reached from state s when performing action a at time t. At each step, the value of a state action pair is updated using the temporal difference term, weighted by a learning rate α_t. Q-Learning is known to converge to an optimal Q function under appropriate conditions [10].

An important condition for convergence is the learning rate. The learning rate controls how much new information is acquired during each cycle. When the learning rate eaches zero, the system has completed its learning. With Q-Learning, the learning rate is modeled as a function.

The learning rate, α, may be controlled in a linear or exponential manner using the parameter w. In the synchronous Q-Learning algorithm, the learning rate decreases exponentially over time and is the same for all state-action pairs:

$$\alpha_t = \frac{1}{(1+t)^w} \text{ and } w \in \left(\frac{1}{2},1\right]$$

In asynchronous Q-learning, the time step, and thus the learning rate, may be different for each state-action pair (s, a).

$$\alpha_t = \frac{1}{(1+t_{s,a})^w} \text{ and } w \in \left(\frac{1}{2},1\right]$$

When rewards are stable over time, asynchronous Q-Learning converges faster than synchronous Q-Learning.

In our experiments with asynchronous Q-Learning, the system appears to forget as soon as it learns. This phenomenon can be explained by observing that humans do not always explicitly reward correct social behavior. Because the learning rate is smaller than unity, without reward, the value of a given state-action pair decreases, effectively causing the system to treat absence of reward as punishment.

We believe that one of the main reasons that learning methods fail in social learning is the use of a learning rate that only depends on time. It is known [19] that for human learning, many other factors, such as social context, time of day, emotional stimulation, motivation, and attention can all have a direct influence on the rate of learning. Thus, we propose to extend the classical learning rate to a multi-dimensional function that depends on different social and environmental factors by creating a rate function $\alpha_t()$ with parameters for state, action, reward, attention, and humor. For convergence purposes, α_t decreases with time under all factors. However, the rate of decrease depends on social and environmental factors.

In the experiments reported below, we have constructed our learning rate function artificially to explore variations of learning rate with situation, action, reward, attention and humor. As long as the system encounters the same situations and actions, the learning rate decreases. We believe that this approach can be extended to other factors by adding additional parameters to the function $\alpha_t()$.

Delayed rewards enable systems to learn which of their actions are desirable based on assigning rewards to actions that occurred in the past. A common approach is to incorporate the use of an eligibility trace [17, 24] to propagate rewards back in time. In our experiments, this approach has not proven useful for social interaction of inappropriate credit assignment.

Rather than reward the most recent state-action pair, we propose to use heuristics to designate the state-action pair that is responsible for the reward (s_e, a_e). Heuristics are modeled as functions that designate a recent state-action pair for reward. The influence of the reward is propagated depending on different factors:

Depending on reward received. People tend to sanction social faults, but rarely reward correct behavior. Furthermore, positive rewards tend to be given for long sequences of action. As with Thomaz [28] we have applied a larger time influence for positive rewards, and a tighter temporal band when for negative rewards.

Transition Probability. Transition probabilities from state-action pair to states are learned at run-time. The influence of a reward is increased with system experience i.e. run-time duration.

Temporal Dependence. Influence decreases as a function of distance from (s_e,a_e). The reward is applied at the designated state action pair as a decreasing function to temporally adjacent state action pairs.

Advice. When in doubt, the system asks which actions were responsible for the reward.

In the experiments below, we use the following eligibility discount function

$$e(s_n, a_n, s_p, a_p, t) = e^{-\frac{1}{2}\left(\frac{t}{\sigma_{stm} \cdot e^{-\frac{1}{2}\left(\frac{1-r}{\sigma_{rwd}}\right)}}\right)^2} P(s_n \mid s_p, a_p)$$

where (s_n, a_n) is the state-action pair to receive the eligibility, (s_p, a_p) is the state-action pair that precedes (s_n, a_n) in the eligibility trace, and t is the time elapsed between (s_n, a_n) and (s_e, a_e). We use σ_{stm} to model the effect of time, and σ_{rwd} to model the effect of the reward on the discount factor. With this approach the learning process has been found to converge in a stable manner.

3.2 Using Analogy to Accelerate Learning

An important advantage of using reinforcement-learning methods is that they allow systems to learn using only scalar feedback. On the other hand, this approach requires that the system perform each state-action pair at least once, and possibly an arbitrarily large number of times. As a result, the convergence time grows exponentially as the number of states and number of actions increase. An important challenge in social learning is to provide algorithms that can enable systems to learn from a few examples without visiting all possible situations and performing all possible actions. Different approaches have been proposed [24, 21] for this problem.

We have explored the use of analogy to accelerate learning. Analogy is a cognitive mechanism that people use to generalize experience to cover similar situations. We propose to modify the Q-Lambda algorithm by adding analogy to propagate information to similar states.

We define analogy using operators. Operators are functions that transform a state into a list of states:

$$op(state) \rightarrow \{anotherState, \ldots\}$$

A state is said to be an analog of another state when it can be derived from this state using an operator. In our approach, the situations are viewed as states defined as a conjunction of relations (truth functions) whose arguments are the entities that have been assigned to a set of roles. Operators can be obtained from simple statements such as "what is true for situation e_1 will be true for situation e_2". For example, a typical operator is "Swap Roles". This operator exchanges the entities playing the roles.

$$SwapRoles([r_1, e_1], [r_2, e_2]) \rightarrow ([r_1, e_2], [r_2, e_1])$$

and thus applies an experience learned with one entity, to all other entities that can play the same role.

We also use operators that provide a list of situations that are similar-to but simpler (have fewer relations) than a given situation. For example RemoveEntity() produces a list of situations in which the relations concerning that entity have been removed.

4. Experimental Evaluation

In this section we describe several experiments with modifications to the classic Q-Lambda learning algorithm, operating in a social environment. In our experiments, we have used scenario-based evaluation. With this approach, we create a scenario involving human actors. Each human follows a rough script and is asked to reward or punish the system each time it acts in a polite or impolite manner. We allow the system to accumulate experience by replaying the scenario repeatedly, while evaluating the results of learning.

4.1 Evaluating Social Learning

Several different methods can be used to evaluate reinforcement learning for social situations.

Cumulative Reward: When the system learns from a scenario, the number of negative rewards should decrease and the number of positive rewards should increase. Thus we can use the cumulative number of negative and positive rewards as a possible measure for evaluating the efficiency of learning.

Frequency of rewards. The frequency of negative or positive rewards over time period gives us information on the rate of change of cumulative reward, and thus reflects the system's current learning rate.

Analysis of the Q-Table: For each state-action pair, the evolution of its Q-Value and indicates how well the Q-Value has converged. Typically when the Q-Value for a state-action pair (s, a) is stable, this indicates that the system has learned the value of the action *a* for the state *s*. This method is used in the first four experiments below.

Analysis of human opinion: As the system should learn to behave socially, it is possible to validate an approach by asking actors for their opinions. Opinions can be obtained by asking actors to complete a questionnaire asking that they rate the system between autistic and sociable on a scale of 1 to 10.

We have selected cumulative reward as a measure of the effectiveness of learning.

4.2 Experimental setup

In our experiments we have used a Sony AIBO robot as a physical embodiment for an interactive system. In particular we have avoided the hard problem of emotion recognition by using the tactile and acoustic sensors available on the AIBO head and back. These sensors have made it possible to provide positive feedback by caressing the head or back sensors and to provide negative feedback by tapping the robot on the head. We have also used a library of pre-programmed speech, song, gesture and dance actions available for the AIBO robot as possible behaviors.

Experiments were performed within the INRIA-Grenoble Smart Environments experimental facility, shown in Figure 1. This facility is an experimental laboratory equipped with furniture for simulating domestic, office and meeting environments, while observing activities with large number of cameras, microphones and other sensors. The facility is constructed with an infrastructure for easily placing cameras and microphones and connecting them to dedicated computing facilities in an adjacent room.

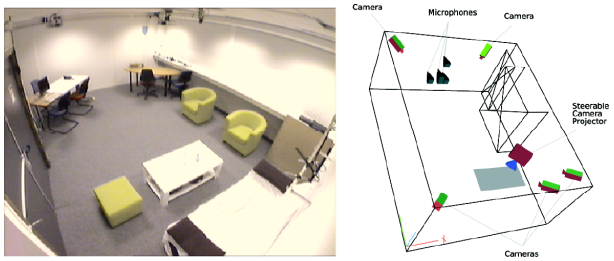

Figure 1. The INRIA Grenoble Smart Environments Facility.

We used a network of cameras and microphones to maintain a situation model describing the activity of one or more humans. In particular, situations were defined using a model of shared

attention proposed by Maisonnasse [18]. The sensors on the AIBO robot were used to drive learning.

For our experiments, we have defined a situation model composed of shared attention of actors, completed by seven activities, associated with positions in the environment. The attentional matrix is used to capture the shared attention between actors. Regions of the room are used to define predicates that correspond to activities.

The dashed circles in figure 2 show six activity regions defined for *working*, *reading, sleeping*, *playing, entering* and *calling on the phone*. A seventh activity, *unknown*, is defined for positions outside these six regions. Formally, each region defines a role that can be played by an actor. The situation is defined as a conjunction of the relations Role-x-is-played-by(actor) (x is one of the seven activities) and attending-to(actor-1, actor-2), where the AIBO is one of the actors.

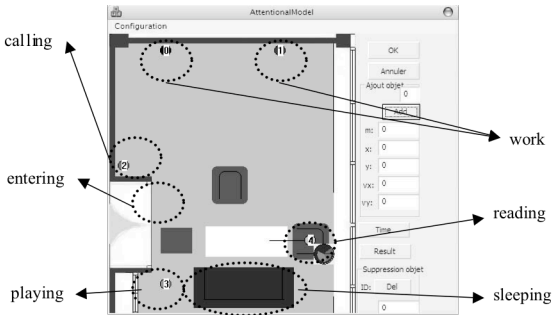

Figure 2. A floor-plan view of the activity regions

For each situation, we allow the AIBO to choose between two actions: bark and play. The first four experiments were based on scenarios defined using a single human actor plus the AIBO. In the fifth experiment, we have added a second human actor to explore reasoning by analogy.

5. Experiments and Results

We have performed five experiments to examine the effectiveness of different forms of reinforcement learning for acquisition of a table of appropriateness for a list of actions. In each experiment, we have analyzed and compared the convergence of the learning algorithms using cumulative reward and the convergence of state-action value functions. The fourth experiment demonstrates the importance of proper credit assignment. The fifth experiment demonstrates the improvement obtained with a modified Q-Learning algorithm using analogy. In this case, the system is able to recover from mistakes in learning and thus to use analogy to reduce learning time. Although the learning algorithms were trained by repeating each scenario multiple times, because of space limitations, we use selected output traces to illustrate properties of each algorithm.

5.1 First Experiment: Standard Q-learning

Our first experiment explored the problems encountered when applying a standard Q-Learning algorithm for social learning. In this experiment we used a single situation (PLAY) in which the actor is in the playing activity region. We allow the AIBO to choose between two actions: bark and play. Actors are asked to divide their time between attending to AIBO or not. When attending to AIBO, actors are asked to give positive feedback (Caress AIBO's back) when AIBO plays and to give negative feedback (tap the AIBO's head) when AIBO barks. When not attending to AIBO, no feedback is given.

A typical result of the first experiment is shown by the three graphs in figure 3. The upper graph shows AIBO's action sequence with gray representing Bark, and black representing Play. The middle graph shows the learned Q-Value for the situation action pairs (PLAY, Bark) in gray and (PLAY, Play) in black. The third graph shows the cumulative reward for this situation.

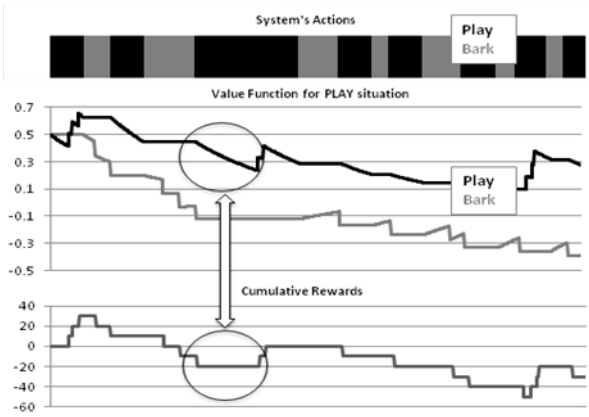

Figure 3. Typical results of the first experiment.

We remark the system did not learn the appropriate actions for the given situation but continued to alternate between the two actions. At the beginning of the experiment, the human actor rewarded AIBO for playing and as a result, the value for this action increases. However, the human actor naturally decreased the reward as AIBO continued to play. Without additional positive feedback, the value for play decreased, and the system forgot the lesson.

This experiment shows that with standard Q-learning, the system requires constant feedback to behave correctly. Unfortunately, rewards given by humans for social actions do not naturally remain constant but depend on different social factors. For example, adult humans do not receive rewards when they brush their teeth (unless perhaps they like the taste of toothpaste), yet continue to apply lessons learned as children. We conclude from this that learning rate must be adapted to fit social constraints.

5.2 Second Experiment: Learning rate

Our second experiment explores the results obtained by modifying the asynchronous Q-Lambda algorithm using a multi-dimensional alpha function, based on the whether the user is attending to AIBO. The experiment takes place in the same condition as in the first experiment, using 2 situations, PLAY and IGNORE. The results of the second experiment are shown in Figure 4, using the same layout of graphs as in figure 3.

Compared to the first experiment, we can observe significant changes. First of all the system correctly learned which actions to perform in each situation. The difference between both Q-Values is significant which means that AIBO learned a preference for the *Play* action in the PLAY situation. The cumulative reward did not become negative which means that the system received more positive rewards than negative ones. We observe that most of the positives rewards are given after the system has been punished and changed its behavior.

Other important changes can be observed. First the system learned faster in the second experiment because the influence of the attention on the learning rate. Indeed we choose that attention

increases the learning rate, as a result influence of reward is much more important in this second experiment. We observe however that the system still forgets when no rewards are given, but does so less rapidly than in the first experiment, because learning rate is much smaller when no feedback and attention is received. We remark as well that the influence of the reward (both negative and positive) grows weaker with time, which guaranties the convergence of our algorithm. The use of a multi-dimensional learning rate function greatly increases the effectiveness of standard Q-Lambda algorithm for learning through social interaction.

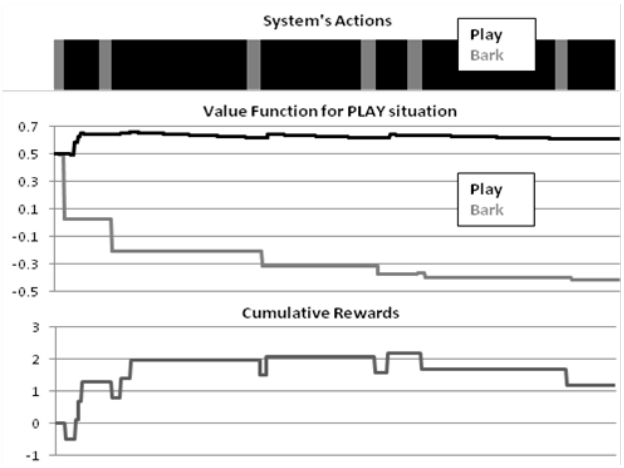

Figure 4. Results of second experiment

5.3 Third Experiment

The third experiment illustrates the inefficiency of the classical Q-Lambda approach when dealing with delayed reward. In this experiment we employed two situations and allowed AIBO to choose between two actions: Play or Sleep. Situation S_{11} is the situation where a person is in the Play region and is paying attention to AIBO. In this configuration, the actor is to give positive rewards to AIBO when it plays and negative rewards when it sleeps. In Situation S_{52}, the actor is *calling on the phone* and not paying attention to AIBO. In this configuration, negative rewards will be given when AIBO Plays, and no rewards are given for sleeping.

AIBO is distant from the phone and in state S_{52}, the actor must leave the phone to give feedback, introducing a temporal delay. We expected AIBO to learn to sleep when a person is on the phone, and to play when the person is in the PLAY situation. However, the results, illustrated in the trace shown in figure 5, were not entirely as expected.

The first two graphs in Figure 5 show typical situation and action transitions that occurred during the experiment. The third and fourth graphs represent respectively Q-Value for the actions Play and Sleep in both situations S_{11} and S_{52}. The last graph represents the cumulative reward. We observe that although the situation S_{11} is the most affected by the human reward, none of its Q-Values have converged. The system did not learn in either situation S_{11} or S_{52} because the delay in receiving feedback caused the reward to be improperly assigned.

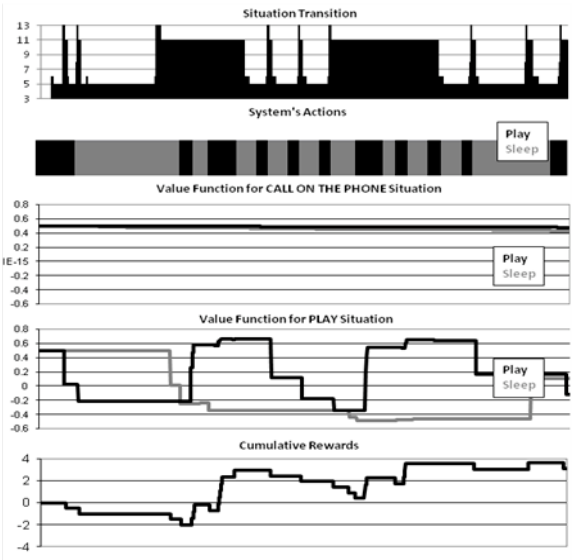

Figure 5. Results from the third experiment

5.4 Fourth Experiment: Delayed Reward

In the fourth experiment we repeated the scenario from the third experiment, with the use of an eligibility trace to assign rewards to situation-action pairs with a time delay. We compared two different credit eligibility traces for credit assignment: Classic (no delay) and delayed with a heuristic to determine the proper action.

Table 1. Two credit assignment techniques

AIBO Historic		Classic RL (Exp 3)		Heuristic (Exp 4)	
Situation	***Action***	***Time***	***Eligibility***	***Time***	***Eligibility***
11	0	0	1.000	5	0.000
13	0	-2	0.003	3	0.000
8	0	-3	0.000	2	0.000
1	0	-3	0.000	2	0.003
6	0	-4	0.000	1	0.100
5	0	-5	0.000	0	1.000
5	1	-12	0.000	-7	0.000
5	1	-20	0.000	-15	0.000

Table 1 shows how propagation of rewards is managed in the eligibility traces, when the human is on the phone and acts to give punishment to AIBO. The first column represents the succession of situation-action pairs observed, with a box for the situation-action pair for which the feedback is intended. The remaining columns illustrate how a reward affects the situation-action pairs for classic reinforcement (as in experiment 3) and delayed reinforcement using a Heuristic to select delay.

Figure 6 shows a typical trace of the results from this experiment. We observe that Q-Values converge for both situations S_{11} and S_{52}. In particular, in the situation S_{52}, AIBO correctly learned that it should sleep while in situation S_{11} it learned to play and not to sleep. We see that with the heuristic, the learning algorithm was

able to correctly find the situation to which the feedback should be assigned while the classical method wrongly assigns the feedback to a later state.

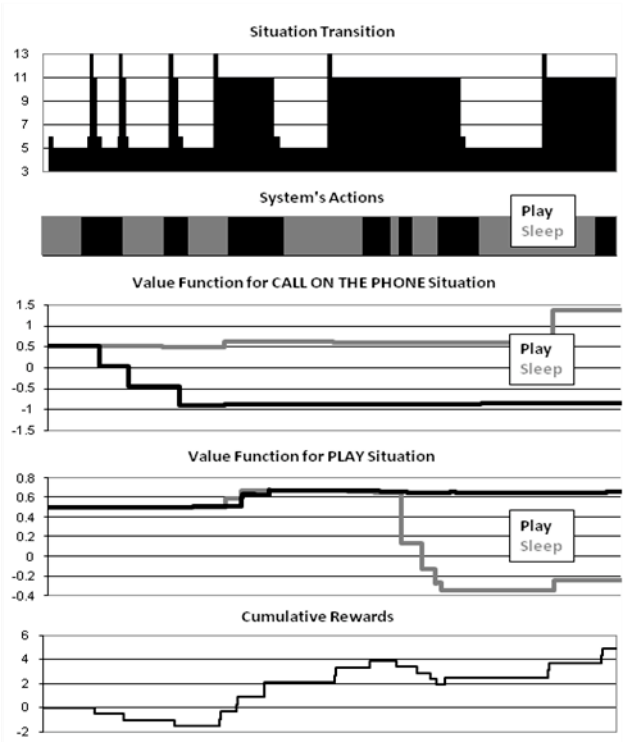

Figure 6. Results from the Fourth experiment

Fifth Experiment: Using Analogy

The firth experiment investigates the use of analogy when learning in a large state space. This experiment has been performed using two actors (E_1 and E_2). We proceed in two phases. In the first phase, E_1 is asked to perform a reading activity while E_2 performs a variety of activities (*entering the room*, *working*, *sleeping*, *reading and calling on the phone*). In the second phase, we invert the roles of E_1 and E_2, so that E_2 reads and E_1 performs different activities. This corresponds to 28 situations.

We allow AIBO to choose between *play*, *sleep* and *say-hello*. Actors were asked to reward or punish AIBO depending on the perceived politeness of its behavior regarding the situation. Here a polite behavior is a behavior that stimulates pleasure while impolite behavior triggers displeasure.

To evaluate the results, we compare the negative vs. the positive rewards obtained in both phases. We remark that during the first phase, the system received more rewards than in the second phase and that a slight majority of these rewards were negative. On the other hand, in the second phase, the system received many more positive rewards than negatives ones.

These result can easily be explained. In the first phase, the system did not have any prior knowledge and thus takes more time to learn to behave correctly. However by using analogy in the second phase, the system has used its past experience with E_2 to choose more appropriate actions for E_1 and thus to obtain more positive rewards. This experiment demonstrates that by using analogy, a system may learn from fewer negative rewards.

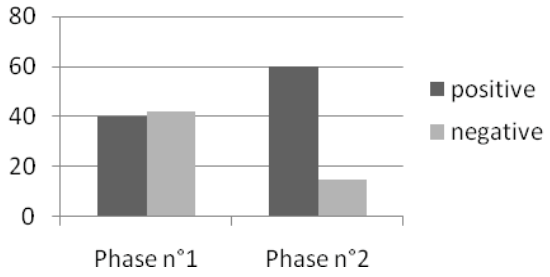

Figure 7. Results from the fifth experiment.

6. Conclusion

Reeves and Nass argue that a social interface may be the truly universal interface [22]. Current systems lack an ability to acquire social common sense because they are unable to learn from social interaction. However, classical reinforcement learning shows a number of weaknesses when applied for learning from social interaction. Our experiments indicate that learning rate, intelligent reward propagation and analogy can each play a significant role in learning social common-sense from social interaction.

For learning rate, the first three experiments have illustrate how social factors can be used to control learning rate to better stabilize learning. We are currently examining the effects of additional social factors such as humor, surprise and anger. Our fourth experiment illustrated that proper credit assignment is necessary for learning for social interaction. Feedback mechanisms that allow designation of the appropriate past situation-action pairs can be useful here. One approach is to allow the system to ask "why", when confronted with feedback for which it is unsure how to properly assign credit. Finally, we have found that analogy can be used to accelerate learning in large state spaces, for example, in contexts composed of many situations. The concept of role, in particular, is useful for both reducing the number of situations, and for defining similar situations for applying analogy.

7. REFERENCES

[1] Adams, B. Breazeal, C. Brooks, R. A., Scassellati, B., "Humanoid robots: a new kind of tool," *Intelligent Systems and Their Applications, IEEE [see also IEEE Intelligent Systems]* , vol.15, no.4, pp.25-31, Jul/Aug 2000

[2] Bartlett, M., Littleworth, G., Fasel, I., and Movellan, J., *Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction*, Workshop on Computer Vision for HCI, CVPR 2003, Vancouver, Canada, 2003.

[3] Brdiczka, O., *Learning Situation Models for Context-Aware Services*, Doctoral Dissertation, INPG, 2007.

[4] Brdiczka, O., Maisonnasse, J., Reignier P., and Crowley, J. L., *Learning individual roles from video in a smart home*, International Conference on Intelligent Environments, 2006.

[5] Breazeal C. and Aryananda, L., Recognition of Affective Communicative Intent in Robot-Directed Speech, Autonomous Robots, 12, 2002.

[6] Breazeal, C., *Designing Sociable Robots*, MIT Press, Cambridge MA, 2002.

[7] Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M., "The Cog Project: Building a Humanoid Robot". In *Computation for metaphors, analogy, and agents*, C. Nehaniv (ed), Lecture notes in artificial intelligence 1562. New York, Springer. 52-87, 1998.

[8] Crowley, J. L., "Context Driven Observation of Human Activity", European Symposium on Ambient Intelligence, Amsterdam, 3-5 November 2003.

[9] De Silva, L. C., and Pei Chi, N., *Bimodal emotion recognition*, FG 2000, Fourth IEEE Conference Automatic Face and Gesture Recognition, pp. 332-335, Grenoble, March 2000.

[10] Even-Dar E. and Mansour, Y., *Learning Rates for Q-Learning*, 14th Annual Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 2001, Proceedings, 2111 (2001), pp. 589-604.

[11] Fong, T., Nourbakhsh I., and Dautenhahn, K., *A Survey of Socially Interactive Robots*, Robotics and Autonomous Systems, 42, 2003.

[12] Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz A. and Wang, J., *Designing robots for long-term social interaction*, IROS 2005, International Conference on Intelligent Robots and Systems, 2005.

[13] Isbell, C. L., Shelton, C. R., Kearns, M., Singh, S., and Stone, P., *A social reinforcement learning agent*, Proceedings of the fifth international conference on Autonomous agents, ACM Press, Montreal, Quebec, Canada, 2001.

[14] Johnson-Laird, P. N. , *How We Reason*. Oxford University Press (2006).

[15] Johnson-Laird, P. N., Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge University Press; Cambridge, MA., 1983.

[16] Kidd, C. D., and Breazeal, C., *Designing a Sociable Robot System for Weight Maintenance*, RO-MAN 2005,14th IEEE International Workshop on Robot and Human Interactive Communication, Nashville TN, Aug 2005.

[17] Klopf, A. H., "Brain function and adaptive systems - A heterostatic theory", Technical Report AFCRL72 -0164, Air Force Cambridge Research Laboratories, Bedford, MA, 1972.

[18] Maisonnasse, J., Gourier, N., Brdiczka O., and Reignier, P., "Attentional Model for Perceiving Social Context in Intelligent Environments", 3rd IFIP Conference on Artificial Intelligence App22lications and Innovations (AIAI), pp171-178, June 2006.

[19] Ormrod, J. E., *Human Learning*, Prentice Hall, 2003.

[20] Padgett, C., and Cottrell, G., *A simple neural network models categorical perception of facial expressions*. In Proceedings of the 20th Annual Conference of the Cognitive Science Society, Lawerence Erlbaum, Hillsdale NJ, 1998.

[21] Preux, P., Propagation of Q-values in Tabular TD(lambda), Proc. 13th European Conference on Machine Learning (ECML), 2430, pp. 369-380, 2002.

[22] Reeves, B. and Nass, C. *The Media Equation: how People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, 1998.

[23] Shin, Y. S., *A Neural Network Model for Classification of Facial Expressions Based on Dimension Model*, *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2005.

[24] Sutton, R. S. "Temporal Credit Assignment in Reinforcement Learning", Ph.D. dissertation, University of Massachusetts, Department of Computer and Information Science, 1984.

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

[26] Thomaz, A. L. and Breazeal, C. *Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance*, Proc. of the 21st National Conference on Artificial Intelligence, AAAI '06, Boston, Mass, Vol 21, Part 1, pp 1000-1005, 2006.

[27] Thomaz, A. L., Hoffman G., and Breazeal, C., *Reinforcement Learning with Human Teachers: Understanding How People Want to Teach Robots*, The 15th IEEE International Symposium on Robot and Human Interactive Communication, pp. 352-357, University of Hertfordshire, Hatfield, Sept 2006.

[28] Thomaz, A. L., "Socially Guided Machine Learning." MIT Ph.D. Thesis, June 2006

[29] Watkins, C. J. C. H., *Learning from Delayed Rewards*, Doctoral Thesis, Cambridge University, 1989.