Proceedings of the IASTED International Conference on AI and Soft Computing, Banff, AB, 2002

Peer Reinforcement in Homogeneous and Heterogeneous Multi-agent Learning

John Anderson[1], Brian Tanner, and Ryan Wegner
Autonomous Agents Laboratory, Department of Computer Science
University of Manitoba, Winnipeg, Manitoba, Canada
andersj,umtanne2,rwegner @cs.umanitoba.ca
http://www.cs.umanitoba.ca/~aalab

Abstract

Reinforcement learning is a broadly employed methodology for training adaptive agents in single- and multi-agent settings. Existing approaches, while being able to vary the type and nature of reinforcement, rely heavily on a centralized omniscient source for reinforcement. This is a significant limitation in terms of modelling human learning: while we do learn directly from skilled teachers, we also learn much from those around us participating in the group activities. Ignoring the latter source of reinforcement severely limits the amount of information an agent can obtain from the world around it. In this paper we explore the use of peer reinforcement – reinforcement obtained from others participating in the same activity, and the effects of employing peer reinforcement for learning in multi-agent systems. We examine two scenarios in a robotic soccer domain to illustrate the use of peer reinforcement in both heterogeneous and homogeneous multi-agent settings.

Keywords: Reinforcement Learning, Multi-Agent Systems, Robotic Soccer, Homogeneity, Heterogeneity.

1. Reinforcement Learning in Multi-Agent Systems

One of the most successful approaches to learning in the control of intelligent agents is reinforcement learning [1]. From the standpoint of intelligent agent control, reinforcement learning restricts the feedback an agent receives as a result of its actions to a positive or negative value reflecting events in the environment (as opposed to receiving sample positive and negative instances of behaviour directly, for example). This creates more work on the part of an agent than example-based learning: an agent must interpret feedback in light of its actions, rather than having them more directly connected by a teacher. This inference can be complicated by the nature of the domain and the nature of the reinforcement. Like learning in general, there is wide breadth of reinforcement learning situations. For example, agents can be reinforced directly action by action, through some synchronous progress indicator, or based on asynchronous environmental events (e.g. a goal scored against one in a soccer game).

While reinforcement learning agents in single-agent settings have significant problems to overcome, each of these problems is exacerbated in a setting where multiple agents interact with one anther [2]. In single-agent reinforcement learning, for example, we must deal with problems of credit-assignment in reinforcement. Agents may obtain reinforcement at a particular time and have to infer the prior actions that could have led to the cause of that reinforcement (the reinforcement is temporally local but must be applied to global series of actions). In individual agents, this is normally handled by propagating local enforcement back to past actions and states [1] in order to reinforce connections between actions. In a multi-agent setting, however, credit-assignment problems are significantly more difficult: not only can reinforcement be delayed, but it may involve the actions of other players as well (an inter-agent credit-assignment problem [2]). Thus, a goal scored in a robotic soccer game (leading to reinforcement) may be due to the actions of a particular agent, in which case those actions should be reinforced. However, it may equally have nothing to do with that agent, in which case any bias based on that reinforcement would be incorrect.

Both Balch [3] and Mataric [4,5] have examined issues of dealing with credit assignment in and between agents in reinforcement learning. Experimenting with temporal locality, Mataric [5] shows that more specialized *shaped* reinforcement, where reinforcement is divided into components and issued as components of a task are completed, improves learning performance. Balch [3] shows similar results, and also illustrates that socially, reinforcement given locally (to an individual agent) provides a similar improvement as a temporally shaped reinforcement over a socially global reinforcement (reinforcement to an entire team without identifying particular individuals). These results confirm intuition: agents improve as we give them more specific information from which to learn, in both individual and multi-agent settings, and form a solid foundation for future work.

One similarity between these works is the source of reinforcement: some (usually omniscient) entity issues

[1] This work is supported by an operating grant from the National Science and Engineering Research Council (NSERC)

reinforcement at a socially local or global level as events in the environment unfold. In some domains, this is largely objective (e.g. a legal goal in soccer). In others however, judgements as to general progress are more subjective (one of the criticisms of shaped reinforcements, because of the attendant dangers of inadvertent experimental manipulation by judges). In either case, we have a single, usually perfect, source of reinforcement.

Human learning goes far beyond this – when we learn in social settings we do indeed get reinforcement from those in authority. In addition to this, however, in social settings we get reinforcement from our peers: when watching a soccer match, or following on-field audio in other sports, one sees and hears a stream of positive and negative reinforcements from individual players to others, and this is clearly something that is often put to a player's advantage.

We are working with agents that learn in multi-agent settings through such *peer reinforcement*. Integrating reinforcement from numerous sources presents difficulties, in that in more complex settings there is the possibility of conflict, and the possibility of having to model the reinforcers. However, peer reinforcement also presents significant opportunities. Consider a soccer match, for example: while it is possible to use natural environmental reinforcements (e.g. scoring a goal), we normally expect human players to learn technique from a coach. While that coach can reinforce players after a game, during time-outs, and (within a limited proximity) directly on the field, the number of these reinforcements is limited (barring the impossible event of having an omniscient coach in direct permanent communication with each agent). Peers, however, can reinforce specific instances of behaviours as they observe them in ongoing behaviour, based on variations of general information supplied before a game by a coach, or based on their own skills. While a peer may not perceive every action on the part of another agent, and not every agent is close enough to every teammate to hear all reinforcements, there are still many more opportunities than a single coach can offer. Through peer reinforcement, we have the ability of reinforcing specific behaviour rather than the result (e.g. the bad behaviour of a goaltender who still luckily managed to stop a ball despite doing the wrong action), *without expecting an ongoing omniscient external reinforcer*. In a robotic domain, peer reinforcement can also assist in correcting misconstrued perceptions, in that others may see and reinforce good situations where a subject agent does not (and by propagating reinforcement back through a history of actions, prior actions can be properly reinforced).

To some degree, peer reinforcement can also be considered a form of shaped reinforcement, such as that seen in [5]. It is often possible to describe the general characteristics of desirable activity beforehand (e.g. a coach telling players to prefer some play, or to stay close to some particularly good player on the opposing team). If players are allowed to reinforce one another based on this general information, the general principles can be translated into specific positive and negative reinforcements as specific situations where the behaviours are applicable arise. We thus arrive at a more specific, context-dependent reinforcement similar in spirit to the temporally shaped reinforcements of Mataric [5].

The remainder of this paper describes an implemented approach to multi-agent reinforcement learning where reinforcement comes from peers as opposed to the traditional omniscient overseer. We examine this approach in both homogeneous and heterogeneous settings in the domain of robotic soccer.

2. Peer Reinforcement in Robotic Soccer Agents

In order to demonstrate peer reinforcement, we have implemented teams of reactive learning agents. These agents use a form of Q-learning to create a mapping between the perception of environmental events and the evocation strength of behaviours. Q-learning [6] involves the use of a function Q, mapping environmental situations an actions to values, to approximate Q*, a mapping to the true value of in the domain of interest. A learning rule is then used to adjust components of Q to better approximate Q* over time. An ε-greedy action selection method is employed to select the best action most of the time, and a random action some small percentage of the time in order to balance exploration and exploitation [6]. A learning rule is used to apportion reinforcement recursively over a number of immediately prior actions.

The implementation domain for this is the RoboSoccer Server [7], a tool for the distributed simulation of robotic soccer games. This is an extremely realistic software simulation for robotic soccer used widely for research as well as for team competition. The RoboSoccer Server provides for visual perception incorporating limitations in terms of both angle and accuracy over distance. In particular, agents are not given direct localization information, and instead must deduce their location on the field by sighting flags marking field landmarks (goal lines, centre field, etc.). The RoboSoccer server also provides for verbal communication between agents (also restricted by distance). We employ the latter facility to allow agents to give reinforcement (positive or negative) including directing that reinforcement to specific agents (the equivalent to shouting "good job player 6" – directed to an individual but perceived by everyone within range).

We are working with peer reinforcement in both homogeneous and heterogeneous situations, with agents that initially know very little about soccer. For a homogeneous situation, we are examining agents reinforcing one another in order to improve basic ball-tracking and movement skills, while for a heterogeneous scenario, we examine players reinforcing a goalkeeper based on the goalkeeper's actions in order to improve the

goalkeeper's defensive movements within the goal crease. In both cases, we are also attempting to employ approaches that are as simple as possible: agents do not have extensive world models, and thus have only a very limited ability to track objects in the world. We have adopted this approach in order to be pragmatic in agent construction as well as to examine this approach to learning under the simplest conditions possible – enhancing the agent's knowledge of the world around them will if anything improve the performance we observe in simple agents.

2.1. Homogeneous Peer Reinforcement

Our homogeneous implementation begins with a largely unskilled team reinforcing one another in specific situations according to general instructions received by a coach beforehand. Agents are told it is a positive step for others to stay close to the ball, and that it is positive for others to kick the ball toward the opponent's goal. We begin with these low level skills for two reasons – they are crucial to developing higher level skills, and beginning with extremely simple agents lessens the possibility that elements already in the agent's behaviour bias learning. While on the field, agents reinforce behaviour in others that they observe, rather than having this reinforcement given objectively (which is not physically possible move by move in a game of soccer). Agents receiving reinforcement may not have seen what is being reinforced (e.g. they may not know they are close to the ball), and may not hear every reinforcement, thus making this similar to a human sensory limitations. Agents have very basic operations (turn in 8 relative directions; kick the ball (on the same 8 directions); or move forward or backward a short (2 units) or long (10 units) distance. These actions are treated independently, resulting in a potential choice of 12 movement actions and 8 kicking actions.

Agents can perceive within a limited distance the relative distance and angle of the ball, as well as that of other players and flags marking specific points on the field. We do not attempt to maintain an accurate position on any of these items, but do attempt to record high-level information about the ball and goal over time. We measure the relative angle of the ball (10 intervals of interest) and the qualitative distance of the ball (4 ranges from near to far). Since the angle of the ball is more crucial than the distance, we attempt to update this angle based on agent rotation over time (distance is not tracked over time). We do however maintain a degree of confidence in the agent's beliefs concerning the ball's direction and distance, since the ball can move rapidly and an agent's action choice should be biased when it knows the rough location of the ball. We start with 100% confidence when the agent actually perceives the ball, and degrade that confidence by a step of 5% each time unit following the perception. We break that confidence down into four ranges (0-30%, 30-60%, 60-90% and >90%) for perception. Adding these confidence intervals to the ball information, we end up with a possible 160 perceptual states. For kicking (only), we need to perceive the direction and distance of the goal as well, in order to know whether a kick deserves peer reinforcement. To keep the size of the table representing our Q function (the Q-table) small, we factor these perceptions into a separate table for kicking actions alone, making the non-kicking Q-table 160 states x 12 actions. Collectively these choices were considered reasonable after experimentation and adjustment over several iterations of this work.

We do not examine perceptual differences over time, and so the agent cannot see which direction a ball is travelling. Instead, we employ a simple heuristic for perception of a goal-oriented kick: an agent informs others (shouts) when it makes a kick, and if others perceive the ball as being between the kicking agent and the goal, a kick toward the goal is assumed. If the ball is on the other side of the agent, a kick away from the goal is assumed. While this does not account for every kick on goal because of complexities in the angle of the perceiving agent, it accounts for most and makes for a simpler agent than one that attempts to analyze perceptual differences over time.

The set of perceptions for kicking decisions includes confidence in the ball direction and goal direction (the same four ranges for each, resulting in 16 possible states). The condition upon whether to kick or move is based entirely on the proximity of the ball, and is not learned.

An agent receives a unit of positive reinforcement from a peer when it is seen to be within 1.5 distance units from the ball or when it is perceived to have made a kick on the opponent's goal. Negative reinforcement is given when a kick away from the opponent's goal is perceived. Reinforcement is carried back through the history of an agent's actions by reducing it 25% for each step in the past, in order to deal with credit assignment over time. Each new reinforcement causes this history to be reset, so a past action can receive at most one reinforcement.

One interesting element in this is that negative reinforcement was not required in order to avoid "bunching up" on the ball. Once learning has progressed so that kicking occurs reasonably often, the ball is in frequent motion, and agents spread out because of this and their limited perception. Even before this, because of the delay between perception and reinforcement, agents receive reinforcement for movements made close to the ball before they were sighted by a peer, as well as for getting to the ball itself, leading to greater agent motion. While negative reinforcement for movement will likely be required in more complex scenarios, the dynamics of the situation suffice to allow agents an adequate chance to move the ball here.

Each Q-table entry records a running average of the reinforcement received by a particular state-action pair, and in response to a perceptual state, the agent selects a random

entry in that row a fraction of the time (ε) and the best action the remainder. ε is varied over time from 90% in the first 8 times a particular perceptual situation is encountered, decreasing to 20% in 10% intervals with each 8 occurrences of that situation.

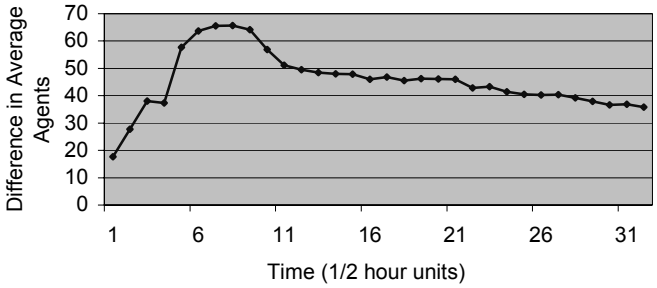

Figure 1. Difference in Q-Tables between average peer-reinforced and perfectly reinforced agents.

To examine peer reinforcement over time, we first required an objectively good agent for comparison purposes. Using the soccer server, we ran a team receiving peer reinforcement against a team of agents that received perfect omniscient reinforcement from a coach for every action (as stated in Section 1, this is completely unrealistic for a real soccer team, but provides an optimum for purposes of comparison). Averaging the Q-tables of all agents on each team and comparing the two teams in terms of the total differences between the squares of their table entries yields the results depicted in Figure 1. This illustrates the peer-reinforced team and the perfectly reinforced team becoming radically different initially, as the team with artificially perfect reinforcement learns faster. However, after the first 4 hours of training (at approximately one action every 25 milliseconds), peer reinforcement begins to reverse this difference and continues to improve over time.

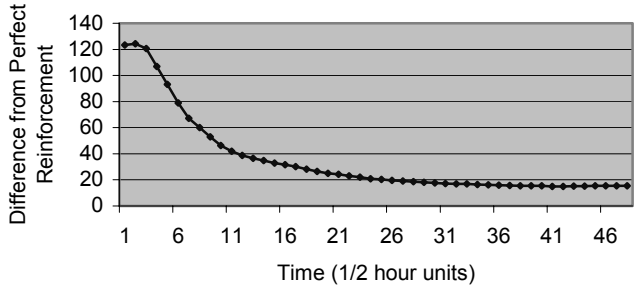

Figure 2. Difference between average of peer-reinforced and trained perfectly reinforced agents over time

After this training, the perfectly reinforced team average was used as an optimum for comparison to examine the performance of peer-reinforced learning over time. Two teams of peer-reinforced agents were set in play against one another, and again the average of all agents was taken and the difference (calculated as above) was examined between the peer-reinforced average and the objective, perfectly reinforced team trained earlier. These results are depicted in Figure 2.

Figure 2 illustrates that over time, a peer-reinforced team will approach the performance of a perfectly reinforced team. The difference when the training exercise was terminated was approximately 15 and was still in flux at that point. We believe continued presence of difference can be accounted for by the fact that after that training there are some significant differences between individual agents, skewing the average. Balch [8] reports similar diversity in learning in soccer, gravitating toward roles based on past experience. We will be running additional trials in future and comparing agent by agent to test this.

These results provide an initial confirmation of the utility of peer-reinforcement: peer-reinforced agents can achieve similar behaviour over time to objectively-reinforced agents. While a team that is receiving perfect reinforcement will learn faster than one that is not, perfect objective reinforcement for each action cannot be expected in most real-world situations.

2.2. Heterogeneous Peer Reinforcement

In the implementation described above, reinforcement comes from peers, but agents are not basing that reinforcement on skills they themselves have mastered – players recognize when others are doing something good or bad and reinforce this, but cannot necessarily do better themselves. While humans do this routinely, in the real world, we do not often see "do as I say, not as I do" reinforcements in so pure a form as that seen in this implementation. This is because in the majority of settings, when learning a task humans have some degree of ability to start with, and if we issue reinforcements to others these include almost unavoidable bias based on those skills. The implementation of the previous Section is more reminiscent of a children's soccer team, where individuals have no skills on which to base their own opinions. While learning in a soccer team made up of human adults would still involve reinforcement based on instructions given earlier by a coach, this reinforcement would be heavily biased (or in some settings replaced) by reinforcement based on that agent's own opinions as to desirable activity.

Once this occurs, we are in a heterogeneous situation, in that a real soccer team will be highly unlikely to have players of identical skills in all respects. It can even occur to some degree in a team that started out as homogenous, due to differences in learning experience. This in fact occurs in the previous experiment – two agents can be significantly different because while they have been running equal lengths of time, one can collect many more reinforcements, especially once it begins learning to kick the ball reasonably well and keeping it moving beyond agents that still track poorly.

Once reinforcements are given that are not purely based on some outside ideal, we have a host of potential agent complexities – from weighing reinforcements against one another to modelling the abilities of other agents in order to weigh reinforcements received from them. While ultimately we are working to deal with all of these issues by experimenting with peer reinforcement in agents that model their teammates (this is the subject of current work), our focus in this paper is on examining the applicability of peer reinforcement in purely reactive agents without regard to higher-level agent functions such as modelling others. In order to both further demonstrate the applicability of this methodology to simple agents and as a step toward dealing with more complex settings, we have examined the use of peer reinforcement in a heterogeneous situation to complement the homogenous work described above.

In the domain of soccer, there is a natural element of heterogeneity in that a goalkeeper has a very different job than any other player on a team. We selected this as a natural choice for heterogeneous experimentation and allowed several non-goalkeeper agents to train a goalkeeper over time. This implementation uses a similar approach to Q-learning as that of Section 2.1. A goalkeeper receives reinforcement from the players training it based on the behaviour those players observe, and gets no reinforcement from external events. That is, the goaltender is primitive enough that if a goal is scored it does not know on its own that this indicates a problem with its behaviour.

In this implementation, a goalkeeper localizes by sighting two field marker flags on the same horizontal and vertical line, then calculating a circle around each with radius equal to the distance of the flag. These two circles will intercept at two points, only one of which will be on the field, and this point is the location of the agent. From this the agent extrapolates its perception of location (we divide the potential locations in the vicinity of the goal into a grid of 9 units) and direction (facing forward, reverse, or neither). The agent also perceives the ball at 7 distance ranges and 4 points of direction relative to itself, yielding a total of 756 perceptual possibilities. It is possible to reduce this by viewing the field as symmetric on either side of a centre line, which is another element of current work.

The actions a goalkeeper can perform are turning toward the ball, or turning to 22, 45, or 67 degrees on either side relative to the angle at which the ball was last perceived (7 turns total), as well as moving forward or backward, or doing nothing. To at the same time examine shaped reinforcement, we also included single action entries involving pairs of actions, coupling each turn with a forward and backward movement, resulting in 14 additional actions and 24 possible learned responses to any perception. Catching is an additional action, which the agent attempts automatically when it is within 1.5 units of the ball and has any confidence in the ball's location (confidence is maintained as in the previous setting, but is used here only for deciding whether to attempt a catch).

A teammate allocates a unit of positive reinforcement (through communication) when the goalkeeper is perceived to be close enough to the ball (one unit of space) to attempt a catch. Similarly, a teammate provides a negative unit of reinforcement when the ball is perceived to be within 1.5 units of the goal line. Reinforcement passes back through the agent's action history in as before, and the change in ε over time is made slightly higher (changing after 30 actions rather than 8) but over time still yields the same balance of exploration to exploitation.

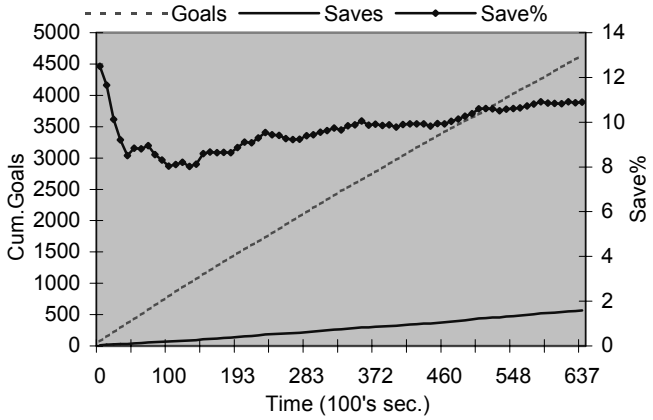

Figure 3. Cumulative goals, saves, and save % as simulation progresses, in perfectly reinforced agents.

To examine the efficacy of peer-reinforcement in this setting, we set a standard by examining learning over time in a team of perfectly reinforced agents (reinforced objectively and accurately upon every goal and save). The results of this are depicted in Figure 3. Cumulative goals (made by 3 trained players) and saves as the simulation progresses are shown according to the left hand scale, and the save percentage, illustrating learning from reinforcement, is plotted on the right hand scale. After some initial outliers as learning begins, the goalkeeper's performance increases steadily as training progresses.

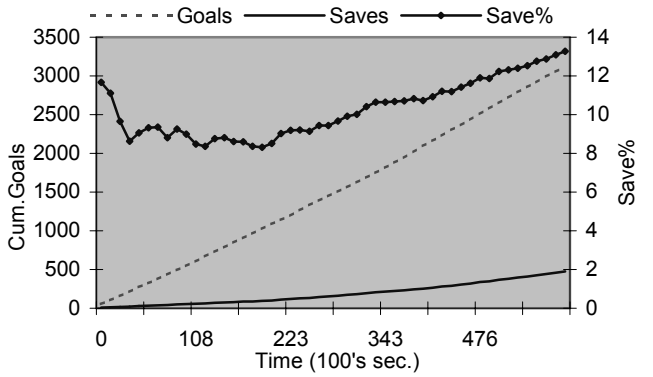

Figure 4. Cumulative goals, saves, and save % as simulation progresses, in peer-reinforced agents.

The results in the identical setting for a team of peer-reinforced agents are illustrated in Figure 4. Here peer reinforcement is delivered by the same agents attempting to score on the goalkeeper, representing a team practice situation as opposed to a real game where one's opponents would not be expected go give such reinforcement. This is a minor consideration however, we could just as easily have the goalkeepers teammates observe goals scored by others. This would, however, have required more players and a slower simulation, resulting in less goalkeeper training in the same overall time.

After the same initial outliers as learning begins, learning in the peer-reinforced setting (the same 3 scoring agents here reinforce the goalkeeper based on their perceptions, as would happen in a team practice setting) actually proceeds with a slightly steeper curve. The performance of this goalkeeper at the end of the simulation was better even though the simulation was stopped before the perfectly reinforced team. While objectively under different combinations of parameters we have found much better goalkeeper performance (on the order of 25-30% saves), this setting does illustrate that the performance of peer reinforcement is just as good as that using objective reinforcement. While the agent can receive multiple reinforcements from peers for the same goals and saves, not all are perceived by the recipient, and not all peers see enough to give a reinforcement in any action episode. The actual increase in performance here we attribute to peer reinforcement as a form of shaped reinforcement: reinforcement is being given in more specific contexts, and this also reinforces the results of Mataric [5] and Balch [3] indicating that such reinforcements can improve learning.

3. Discussion and Future Work

In this paper we have demonstrated the efficacy of peer-reinforcement in multi-agent learning, using simple reactive agents. While peer reinforcement in a homogeneous situation takes longer to achieve the same results, this form of reinforcement is much more realistic than the objective reinforcement at all times assumed in other approaches. In the particular heterogeneous situation used here, the shaped nature of peer reinforcement was also evident.

We believe this to be an important form of reinforcement that despite not being covered in modern reviews of reinforcement learning (e.g. [9]), is well-worth future experimentation. We are currently working on extending the reactive agents used here to employ very simple models of other agents, in order to be able to gauge conflicting reinforcements. We will then be able to experiment with this in more subtle areas and deal with other complex issues such as deception in this form of reinforcement.

We are also interested in combining peer reinforcement with imitation. Imitation by viewing or communicating agents' behaviours and the reinforcements they receive for these [4,5, and more primitively in 10, 11]) has also been shown to be of great use in multi-agent settings. We intend ultimately to experiment with combining both the imitation of others and the integration of reinforcements received from them, under both honest and deceptive settings.

Bibliography

[1] Kaelbling, Leslie, Michael L. Littman, and Andrew W. Moore, "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research* 4 (1996):237-285.

[2] Weiss, Gerhard, *Multiagent Systems* (Cambridge, MA: MIT Press), 1999. 643 pp.

[3] Balch, Tucker, "Reward and Diversity in Multirobot Foraging", in *Proceedings of the IJCAI-99 Workshop on Agents Learning About, From and With other Agents*, Sweden, August 1999. 7 pp.

[4] Mataric, Maja J., "Learning Social Behavior", *Robotics and Autonomous Systems* 20(1997):191-204.

[5] Mataric, Maja J., "Reinforcement Learning in the Multi-Robot Domain", *Autonomous Robots*, 4(1, Mar 1997):73-83.

[6] Sutton, R., and A. Barto, *Reinforcement Learning* (Cambridge, MA:MIT Press), 1998. 322 pp.

[7] Noda, I., H. Matsubara, K. Hiraki and I. Frank. "Soccer Server: a Tool for Research on Multi-Agent Systems", *Applied Artificial Intelligence* 12(2-3, 1998):233-250.

[8] Balch, Tucker, *Behavioural Diversity in Learning Robot Teams*, Ph.D. Dissertation, Computer Science, Georgia Tech, 1998.

[9] Stone, Peter, and Manuela Veloso, "Multiagent Systems: a Survey from a Machine Learning Perspective", *Autonomous Robots* 8(3, July 2000).

[10] Andou, Tomohito, "Refinement of Soccer Agents' Positions Using Reinforcement Learning", in Kitano, Hiroaki (Ed.), *RoboCup-97* (Berlin: Springer-Verlag), 1998, pp. 373-388.

[11] Tan, Ming, "Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents", *Proceedings of the 10th International Conference on Machine Learning*, 1993, pp. 330-337.