In: Proc. BENELEARN 2000, 10th Belgian-Dutch Conference on Machine Learning, 2000

The influence of communication on the choice to behave cooperativly

K. H. W. J. ten Tusscher*, S. H. G. ten Hagen†, M. A. Wiering‡

* Utrecht University, Faculty of Biology
Padualaan 8, 3584 CM Utrecht
† University of Amsterdam, Faculty of Computer Science
Kruislaan 403, 1098 SJ Amsterdam
‡ Utrecht University, Faculty of Computer Science
Padualaan 14, 3584 CM Utrecht

Abstract

In this paper we investigate the learning of cooperation and communication in a multi agent system. A predator prey pursuit domain is defined in which predators can learn to both non-cooperatively and cooperatively capture prey. We then study the influence of communication on a predator's choice to cooperate. Communication will be learned based on its enhancement of cooperation. We show that the developed communicative abilities allow the predators to make a more optimal choice between the cooperative and non-cooperative behaviour.

1 Introduction

Cooperation is defined as the coordinated acting of agents that are part of a multi agent system. Since agents are supposedly selfish, cooperation is thought to occur only if pay offs are involved for all cooperating agents, be it immediately and personally or on the long term and through closely related individuals [Grim, 1996, Brauchli et al., 1999, Cohen et al., 1999, Oliphant, 1994]. Communicating is most often seen as the exchange of information between a sender and receiver of an emitted signal [De Jong, 2000a, Aitchison, 1997].

It has therefore often been suggested that communication can improve cooperation by serving as a means to coordinate the behaviour of agents [De Jong, 1997, 2000b,a, Tan, 1993, Rooijmans, 2000, Matarić, 1998]. Research demonstrating that this indeed is the case is often limited to cooperative tasks for which the communication is necessary, while the communication often is prewired [De Jong, 1997, 2000b, Matarić, 1997].

In research investigating the origins of communication it is often assumed that the communication is advantageous on its own. In other words, the goal of communication is to understand and be understood [Steels, 1996, 1997, De Jong, 2000a]. Under this assumption, communication can be shown to arise through learning or evolution in groups of interacting agents.

We felt that it would be more realistic to combine the two above mentioned approaches when studying the influence of communication on cooperation. We therefore focussed on the contribution of a learned communication system to a cooperative task that can be performed considerably well without communication. The learning of the communication is coupled to the contribution it makes to cooperation. Communication will thus only arise if it is able to contribute to cooperation.

In addition we will not investigate the influence of communication on a stand alone cooperative task, but instead study the influence of communication on the choice to behave cooperatively or not. We hypothesise that if communication enhances cooperation, it will also bias the choice between a cooperative and alternative, non-cooperative behaviour towards the cooperative behaviour.

In section 2 we will discuss the set up of our model and the used learning algorithm. Section 3 describes the architecture of our agents. In section 4 we formulate a hypothesis about the extent to which communication is expected to influence the choice between cooperative and non-cooperative behaviour. In section 5 outcomes of the model are demonstrated. In the discussion of section 6 these results are explained and in section 7 general conclusions are drawn. Section 8 concludes with some proposed future research.

2 Model set up

2.1 Problem domain

As a problem domain we take a predator prey pursuit domain. The problem domain is relatively simple and has already been used by others to study the influence of communication on cooperation [Tan, 1993, De Jong, 1997, 2000b].

In general terms a predator prey pursuit domain consists of a world inhabited by predators and prey in which the former should try and catch the latter. We used a discretised grid world of size 20 times 20. Torus boundary conditions (joining of left and right borders and upper and lower borders) are implemented to avoid boundary effects to interfere with the learning process .

The prey move around in a random fashion and can therefore be considered as part of the world. To offer the predators a choice between behaving cooperatively or not, the domain is inhabited by small prey that can be captured by a single predator standing next to it and large prey that need to be captured by two cooperating predators standing next to it.

2.2 Agents

The predators have a square local perceptual field that allows them to perceive the world around them. To limit the number of possible input states, situations requiring the predators to perform the same or similar actions are mapped to a single input state (state aggregation). Here generalisation is achieved by subdividing the perceptual field of the predators into four partially overlapping square tiles spanning an area from the position occupied by the predator to one of the corners of it's perceptual field. Instead of distinguishing per single grid position which agents are present, we do so per tile. As a consequence, situations in which prey or colleague predators are positioned in a similar direction but at different distances are mapped to a single input state (see figure 1). For a precise description of the derivation of the different possible input states we refer to Ten Tusscher [2000].

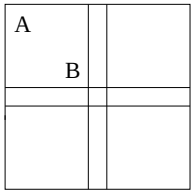


Figure1. State aggregation: by subdividing the perceptual field into four equally large partially overlapping tiles situations in which a prey or predator occupies position A and situations in which a similar type of agent occupies position B are mapped to the same input state.

The action set of the predators consists of making a move to the North, South, East or West. In addition, communicating predators can also emit a signal '0' or '1'. Succesfull behaviour, that is behaviour that leads to the capturing of a prey, is rewarded.

To allow the predators to learn to map their inputs to actions in such a way that they capture as much prey as possible Q-learning [Watkins, 1989, Watkins and Dayan, 1992] is used.

2.3 Q-learning

The basic idea of Q-learning is to define the quality of a state action pair as the expected sum of rewards received when performing an action in a state and apply the policy (input action mapping) from then on:

$$Q^{\pi}(s,a) = \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma V^{\pi}(s')) \quad (1)$$

where

s: current state
s': next state
a: action
π: policy
γ: discount factor
T: state transition probability function
R: expectancy value of reward
$Q^{\pi}(s,a)$: state action quality value
$V^{\pi}(s) = \max_a Q^{\pi}(s,a)$: state value

We would like a fast solution to be preferred over a slow one. In other words we would like the predators to prefer an action leading to the immediate capturing of a prey over an action requiring five more actions before a prey can be captured. To achieve this a discounted sum of expected future rewards is optimised.

In typical applications of Q-learning we do not know R but only know the actually received reward r. In addition, we do not know T but only the actually occurring next state s'. By exploring different possible actions and experiencing feedback from the environment, Q-values can be approximated.

It can be proven that by using the following update rule and decreasing the learning parameter α appropriately, the weighted sum of values for the different possible next states and rewards is computed

accurately:

$$Q^{\pi}(s,a) = Q^{\pi}(s,a) + \alpha(r + \gamma V^{\pi}(s^{'}) - Q^{\pi}(s,a)) \quad (2)$$

The Q-values can then be used to select the optimal action: the action with the highest Q-value. The optimal actions for all possible input states together constitute the optimal policy.

3 Agent architecture

3.1 Non-communicating predators

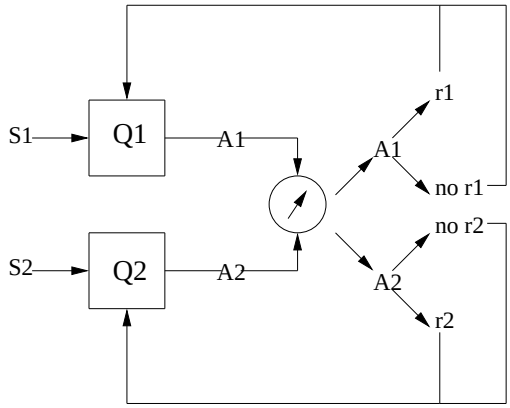


Figure2. Architecture of non-communicating predators. Meaning of the used symbols: S: input state, Q: Q table or expert, A: action and r: reward.

In figure 2 the architecture of non-communicating predators is depicted. A non-communicating predator consists of two separate experts or Q-tables, responsible for learning the two separate tasks. The expert responsible for capturing small prey (Q1) observes whether small prey (S1) are present, while the expert responsible for capturing large prey (Q2) sees whether large prey and colleague predators (S2), both needed for efficiently capturing large prey, are present [Tan, 1993].

The expert with the highest variance in Q-values decides which move is made. The idea behind this is as follows: imagine that one of the experts has Q-values that are approximately equal for all possible actions while the other expert has Q-values that are widely different for the different possible actions. Obviously, for the first expert it hardly makes a difference which action is taken, while for the second expert some actions would be far better to perform than others. By letting the expert with the highest variance decide, the second expert gets to decide what to do. The variance thus more or less reflects how important it is for a particular expert to get to decide which move is made. It should however be noted that this is true once a task has been accurately learned, but that this is not necessarily the case during learning.

The expert being in charge is updated according to the reward received for the particular task that expert is concerned with (r1, no r1 or r2, no r2). In other words, if a predator "intended" to capture a large prey, but "accidentally" captured a small prey, the thus received reward will not be used as an erroneous signal to update the large prey expert. This is done to avoid interference occurring between the learning of the two different tasks.

3.2 Communicating predators

For communication to enhance cooperation, the information that is being communicated should be relevant for the cooperative task and should not be available to the predators without communication. We choose to let our predators communicate about the presence of large prey: a predator not perceiving a large prey signals '0' while a predator that does perceive a large prey signals '1'.

However, for the communication to be of any influence on the cooperative task, the predators have to be able to interpret each others signalling behaviour and use it for their decision which move is to be made. This requires an extension of the predator architecture described in the previous section.

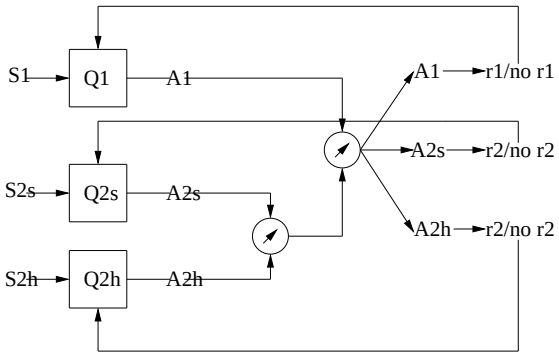


Figure3. Architecture of communicating predators. Meaning of the used symbols: S: input state, Q: Q table or expert, A: action and r: reward.

Figure 3 shows the architecture of communicating predators. For the task of capturing small prey there still is a single visual expert (Q1) perceiving small prey. For the task of capturing large prey there now are two experts, a visual expert seeing both large prey and colleague predators (Q2s) that corresponds to the

single expert present in non-communicating predators (Q2) and an auditive expert hearing colleagues signalling '0' and colleagues signalling '1' (Q2h). The highest variance method is here used first to decide whether the visual or auditive information is most relevant for the task of capturing large prey and then to decide whether the predator is going after small or large prey.

4 Hypothesised contribution of communication

As the visual and auditive perceptual fields are equally large, predators see and hear their colleague predators at the same time. For the large prey predators communicate about, the effective perceptual field gets enlarged: once a predator has learned how to interpret his colleagues' signalling behaviour it can deduce the presence of large prey that he can not see but that is seen and communicated about by his nearby colleague. Communication can thus provide for extra information.

In most situations, visual information will be sufficient to make a well informed choice between the cooperative and non-cooperative task. If however a small prey and a colleague predator are perceived it is hard to decide what to do based on only visual information. In this kind of situations the extra auditive information becomes relevant. It then might be best to go after the small prey if the predator signals '0' (it does not perceive a large prey), while it is best to team up with the predator if it signals '1' (it perceives a large prey).

From this it follows that the contribution of communication to the quality of and choice for cooperation will be limited to a subset of situations in which communication is able to provide for extra relevant information. The question thus becomes whether communication contributes enough for it to be learned and used and perhaps even influence the choice made between behaving cooperatively or not.

5 Experiments

5.1 Model settings

In order for communication to contribute to cooperation at all, circumstances need to be such that the communication indeed provides for relevant extra information. If the grid world is crowded with large prey and all predators see one, communicating about the presence of large prey is not very informative. If the density of large prey however is such that approximately half of the predators perceive a large prey and half of them does not, communicating about the presence of large prey becomes meaningful.

To make the circumstances in our model fit the latter criteria, the grid world of size 20 times 20 will be inhabited by 8 predators, 4 large prey and, for comparison reasons, also 4 small prey. To compensate for the fact that the capturing of large prey is more complex and therefore harder to learn, the large prey reward is made 5 times as large as the small prey reward.

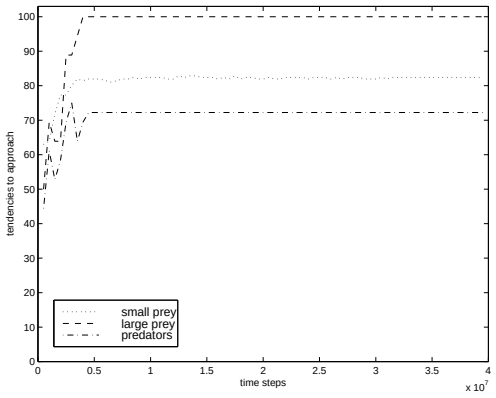
5.2 Non-communicating predators

First we checked whether the non-communicating predators are able to learn to capture both types of prey and learn to chose between these two tasks.

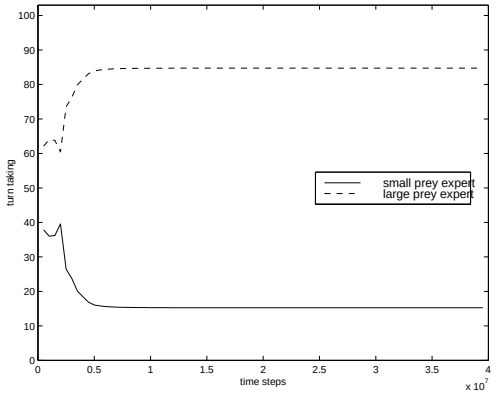
Because of the random movement of the prey, it is not very easy to evaluate the policies learned by the predators based on the number of captured prey per unit of time. Therefore, we had to evaluate the learned policies differently. Because of the relative simplicity of the two tasks it is easy to define what are necessary prerequisites for the tasks to be accomplished. To capture a small prey, a predator has to be able to approach a small prey. To cooperatively capture a large prey, a predator has to be able to approach both large prey and colleague predators. The percentage of situations in which the small prey expert approaches small prey and the tendency of the large prey expert to approach large prey and colleague predators can thus be taken as a quality measure of the learned behaviour.

It should be kept in mind that these tendencies reflect the quality of a particular expert at performing a particular task, but that they do not reflect how often a task actually is performed. To gain insight into the latter aspect, we measure the percentage of situations in which a particular expert decides which move is to be made. Only situations in which both experts receive input are considered. The turn taking of the two experts then reflects a predator's preference for the two behaviours in situations in which either one of them could be performed.

In figure 4(a) the development of the tendencies to approach the different types of agents during the process of learning are shown. The tendencies can be seen to stabilise after an initial period of increase. The small prey expert approaches small prey in approximately 83 percent of the situations, whereas the large prey expert approach colleague predators in 73 percent of the situations and large prey in 98 percent



(a) Tendencies developed by non-communicating predators.



(b) Decision behaviour developed by non-communicating predators.

Figure4. Tendencies and decision behaviour developed by non-communicating predators. Results are averaged over multiple, seperately learning predators.

of the situations. The preference developed for large prey over colleague predators can be explained as follows: to capture a large prey both a large prey and a colleague predator are needed. As colleagues are more numerous than large prey it is more likely for a predator to bump into a colleague once it has tracked down a large prey than the other way around. Approaching large prey thus has priority over approaching colleagues.

In figure 4(b) the development of the turn taking of the different experts during the learning process is shown. For non-communicating predators the turn taking of the visual small and large prey expert reflect a predators preference for behaving non-cooperatively or cooperatively, respectively. The figure shows that approximately 85 percent of the decisions are made by the visual large prey expert. The predators thus show a clear preference for the cooperative behaviour.

5.3 Communicating predators

Now we will study whether communicating predators are able to learn both tasks and learn to choose between them. We also investigate whether the predators learn to use the communicated information to enhance their abilities to capture large prey and make a better informed choice between the non-cooperative and cooperative behaviour.

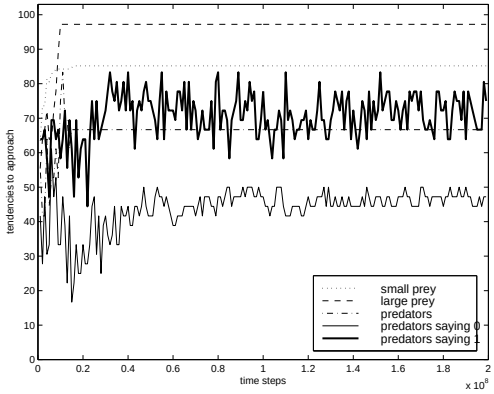
To see whether the predators learn to distinguish between predators signalling '0' and predators signalling '1' and hence learn to interpret each others signalling behaviour, we measure the auditive large prey experts' tendency to approach colleagues signalling '0' and colleagues signalling '1'. In addition we measure how often this expert is in charge to decide which move is made.

Figure 5(a) shows the development of the tendencies to approach the different categories of agents. The tendencies to approach small prey, large prey and colleague predators develop similar to the ones in non-communicating predators. In addition the predators develop a preference for colleagues signalling '1' over colleagues signalling '0'. From the figure it becomes clear that these latter tendencies develop on a far slower time scale than the tendencies to approach small prey, large prey and colleagues.

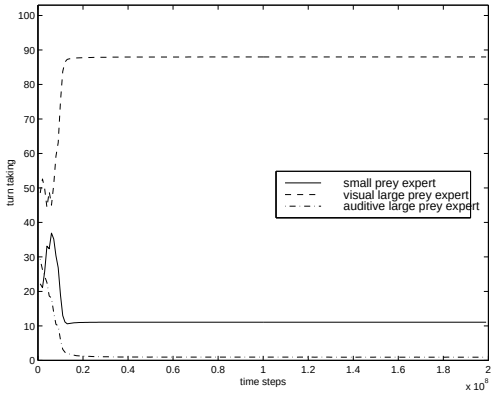
Figure 5(b) shows the development of the turn taking of the different experts. For the communicating predators, the visual and auditive large prey expert together are responsible for the decision to cooperate. From figure 5(b) it follows that the auditive large prey expert makes only a minor contribution to the decision to behave cooperatively. The visual large prey expert decides in approximately 88 percent of the situations. So not only does the extra auditive large prey expert make a contribution to cooperation, also the visual large prey expert makes a larger contribution to cooperation than was the case without communication.

5.4 The influence of communication

In the previous two sections predators could be seen to develop a preference for cooperative behaviour for a situation in which cooperatively capturing a large prey was 5 times more rewarding than capturing a small prey. Here we will study in more detail the



(a) Tendencies developed by communicating predators.



(b) Decision behaviour developed by communicating predators.

Figure5. Tendencies and decision behaviour developed by communicating predators. Results are averaged over multiple, seperately learning predators.

dependence of predators preferences on the relative rewards received for the capturing of the two prey types.

We define the reward ratio as the large prey reward divided by the small prey reward. Given the fact that capturing large prey is more complex than capturing small prey we expect a predators preference to switch from the non-cooperative to the cooperative behaviour for a reward ratio larger than 1. We will furthermore study the influence of communication on the relationship between preference and reward ratio.

We plotted the percentage of situations in which the different experts are in charge for both non-communicating and communicating predators as a function of the reward ratios (see figure 6). These turn taking percentages are the final outcomes of the learning process.

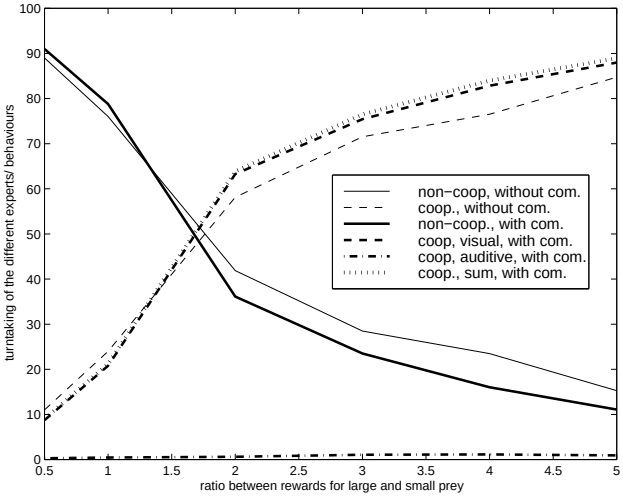


Figure6. Comparison of the choice made by the predators to behave cooperatively or not for predators with and without the ability to communicate.

In non-communicating predators the visual small prey expert is responsible for the non-cooperative behaviour (non-coop., without com.) and the visual large prey expert is responsible for the cooperative behaviour (coop., without com.). The dominance in turn taking of these two experts switches for a reward ratio of 1.9.

In communicating predators a visual and auditive large prey expert are together responsible for the choice to behave cooperatively (coop., visual, with com and coop., auditive, with com respectively). To allow a comparison between the preference of non-communicating and communicating predators for cooperative the turn taking of these experts are summed (coop., sum, with com.).

The auditive large prey expert makes a minor and approximately constant contribution to the decision to behave cooperatively in communicating experts. The dominance in turn taking switches from the visual small prey to the visual large prey expert, similar to what can be observed for non-communicating predators. There are however three differences between the turn taking of the visual small and large prey experts in communicating and non-communicating predators. First of all, for low large prey rewards, communicating predators choose a bit more often to not cooperate. Second, for high large prey rewards, communicating predators choose a bit more often to cooperate. Third, with communication

the switch from predominantly non-cooperative behaviour to predominantly cooperative behaviour occurs for a somewhat lower large prey reward.

6 Discussion

The non-communicating and communicating predators show comparable increases in their tendencies to approach small prey, large prey and colleague predators. This implies that the two types of predators have learned equally well to capture small and large prey based on visual information.

In addition the predators augmented with communicative abilities develop a preference for colleagues signalling '1' over colleagues signalling '0'. This implies that predators learn to interpret each other signalling behaviour. From the turn taking of the different experts it follows that the communicated information is used in only a small set of rare situations where this extra information can help to resolve a dilemma such as whether a colleague or a small prey should be approached (see section 4). This also explains why the tendencies to approach colleagues signalling '0' and colleagues signalling '1' develop on a much slower time scale.

In spite of the small contribution made to cooperation by the auditive expert, communication does influence the turn taking of the cooperative and non-cooperative behaviour. This influence can be understood as follows: by incorporating communication certain complicated situations are left to be resolved by the auditive large prey expert. As a consequence, not only can a better informed decision be taken in those situations, it also allows the visual expert to be trained exclusively on situations in which visual information is sufficient. This causes the visual expert to experience more consistent feedback form the environment allowing it to function more accurately. As a consequence the turn taking between visual small prey expert and visual large prey expert is divided more optimally.

7 Conclusions

We have shown that predators can learn to capture both small and large prey using only visual information. We have also shown that communication is learned and used despite the fact that it only provides for extra information relevant for communication in a limited number of situations. Communication can thus be shown to arise not only if it is assumed to be advantageous on its own but also if it provides for only a slight improvement of a particular behaviour.

We have furthermore shown that communication allows for a better informed choice between cooperative and non-cooperative behaviour. By taking over control in a limited number of situations where it can provide for extra information the auditive large prey expert not only allows for a better informed choice in those situations, but also allows the visual large prey expert to specialise on situations in which visual information is sufficient. The more accurate functioning of the visual large prey expert results in less cooperation if cooperation pay offs are low and more cooperation if cooperation pay offs are high. Communication can thus be shown to provide extra information relevant for the decision whether or not to cooperate in a subset of the occurring situations.

8 Future work

One possible extension of our model would be to vary the number of predators needed to capture a large prey. By increasing the number of predators needed for succesfull cooperation, cooperation is made more complex. It would be interesting to investigate whether communication then can play a more important role in coordinating cooperation and optimising the choice between capturing large and small prey.

Another possible extension would be to let the predators learn both the signalling and the interpretation of each others signalling behaviour instead of only the latter. Because of the restricted contribution communication can make to cooperation, we expect this complete learning of the communication to take very long. A solution to this problem would be to choose a task to which communication can make a more important contribution. This would allow communication to be learned faster.

References

J. Aitchison. *De sprekende aap, translated from: The Seeds of Speech, Language Origin and Evolution.* Het Spectrum, 1997.

K. Brauchli, T. Killingback, and M. Doebeli. Evolution of Cooperation in Spatially Structured Populations. *Journal of Theoretical Biology*, 1999.

M. D. Cohen, R. L. Riolo, and R. Axelrod. The Emergence of Social Organization in the Prisoner's Dilemma: How Context Preservation and other

Factors Promote Cooperation. Technical report, Santa Fe Institute, 1999.

E. De Jong. Multi-Agent Coordination by Communication of Evaluations. In M. Boman and W. Van de Velde, editors, *Proceedings of the 8th European Workshop on Modelling Autonomous Agents in A Multi-Agent World MAAMA*. Springer-Verlag, 1997.

E. De Jong. *Autonomous formation of concepts and communication*. PhD thesis, Vrije Universiteit Brussel, 2000a.

E. De Jong. Coordination Developed by Learning from Evaluations. In *Collaboration between Human and Artificial Societies*, volume 1624. Springer-Verlag, 2000b.

P. Grim. Spatialization and greater generosity in the stochastic Prisoner's Dilemma. *Biosystems*, 1996.

M. J. Matarić. Learning Social Behaviours. *Robotics and Autonomous Systems*, 1997.

M. J. Matarić. Using communication to reduce locality in Distributed Multi-Agent Learning. *Journal of Experimental and Theoretical Artificial Intelligence, special issue on Learning in DAI Sytems*, 1998.

M. Oliphant. Evolving Cooperation in the Non-Iterated Prisoner's Dilemma: The Importance of Spatial Organization. In R. Brooks and P. Maes, editors, *Proceedings of the Fourth Artificial Life Workshop*. MIT Press, 1994.

N. Rooijmans. Coöperatie in Multi Agent Systemen, communicatie en cöoperatie in een systeem van meerdere Q-learning agents. Master's thesis, Utrecht University, 2000.

L. Steels. *Machine Intelligence 15*, chapter The Spontaneous Self-organization of an Adaptive Language. Oxford University Press, 1996.

L. Steels. Language learning and language contact. In W. Daelemans, editor, *Proceedings of the workshop on Empirical Approaches to Language Acquisition*, 1997.

M. Tan. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In *Workshop on Reinforcement Learning, 10th International Conference on Machine Learning*, pages 102–112, 1993.

K. H. W. J. Ten Tusscher. Learning to cooperate by using communication in a simulated predator prey pursuit domain. Master's thesis, Utrecht University/University of Amsterdam, 2000.

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, 1989.

C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 1992.