Learning to Behave Socially

Maja J Mataric
MIT Artificial Intelligence Laboratory
545 Technology Square #721
Cambridge, MA 02139
phone: (617) 253-8839
fax: (617) 253-0039
email: maja@ai.mit.edu

April 10, 1994

Abstract

Our previous work introduced a methodology for synthesizing and analyzing *basic behaviors* which served as a substrate for generating a large repertoire of higher–level group interactions (Matarić 1992, Matarić 1993). In this paper we describe how, given the substrate, agents can learn to behave *socially*, i.e. to maximize average individual by maximizing collective benefit. While this is a well–defined problem for rational agents, it is difficult to learn in situated domains. We describe three sources of reinforcement and show their necessity for learning non–greedy social rules. The learning strategy is demonstrated on a group of physical mobile robots learning to yield and share information in a foraging task.

1 Introduction

Our previous work focused on analyzing and synthesizing complex group behaviors from simple social interactions between individuals (Matarić 1992, Matarić 1993). We introduced a methodology which involved designing a collection of *basic behaviors* which served as a substrate for a large repertoire of higher–level interactions. In this paper we describe how agents can extend the basic repertoire in order to learn to behave *socially*.

The importance of social behavior needs no elaboration. Social rules, from the simplest traffic protocols to the most complex ethical constructs, minimize conflict and optimize global efficiency. Consequently, social rules are a natural part of any society, and the acquisition of those rules is an important aspect of adaptive behavior. This paper will describe how social learning, a fundamental form of adaptive group behavior, can be used for acquiring social rules.

1.1 Social Learning

Social learning is the process of acquiring new behavior patterns in a social context, by learning from conspecifics. Also called *observational learning*, social learning is ubiquitous in nature (McFarland 1985, McFarland 1987) and the propensity for it appears to be innate. Animals imprint, mimic, and imitate adults of their own kind instinctively, often without obtaining direct rewards or even successfully achieving the goal of the behavior (McFarland 1985, Gould 1982). A crucial form of learning during development, social learning remains an important aspect of social interaction throughout an animal's life.

Observational learning is a phylogenetically old method of adapting behavior. It has been found in some invertebrates (Fiority & Scotto 1992), in many birds (Moore 1992), aquatic mammals (Mitchell 1987) and of course primates and humans. In all cases, it is a faster and more efficient form of acquiring new behaviors than its traditional classical conditioning counterpart. Imitation of conspecifics is easier than general imitation and mimicry, but animals, and some birds (Moore 1992), are capable of both.

Social learning includes learning *how* to perform a behavior, through imitation, and *when* to perform it, through social facilitation. *Imitation* is defined as the ability to observe and repeat the behavior of another animal or agent. It is one of the principal modes of acquiring new patterns of behavior. True imitation is distinct from mimicry and social facilitation. *Mimicry* is the ability to repeat some aspects of the behavior of another agent, without having the imitator understand the goal of the behavior or of the internal state of the agent being imitated (Tomasello, Kruger & Rather 1992).

Social facilitation refers to the process of selectively expressing a behavior which is already a part of the animal's species–specific repertoire. Birds pecking at milk bottle caps in order to drink the milk inside (Davis

1973, McFarland 1985), and eating more in groups then when alone (McFarland 1987) are both examples of social facilitation in which the animals are not learning any new behavior but only responding to a releaser (Davis 1973).

Some previous work has addressed the problem of learning behavior selection (e.g. (Maes & Brooks 1990)) and we have also demonstrated an algorithm for learning higher–level group behaviors by selecting among basic primitives (Matarić 1994). Unlike previous work, this paper focuses on learning social rules, i.e. behaviors that do not produce immediate payoff for the agent but benefit the group as a whole. Game theory has dealt extensively with related problems (for example see Axelrod (1984), Kraus (1993) and Durfee, Lee & Gmytrasiewicz (1993)) but exclusively in the domains with rational agents capable of correctly evaluating the utility of their actions and strategies. In contrast, we are interested in finding out what is required for learning social strategies in situated agent domains where, due to incomplete or nonexistent world models, inconsistent reinforcement, noise and uncertainty, the agents cannot be assumed to be rational. In general, systems treated by game theory are usually simpler and more cleanly constrained than those found in biology and thus those that motivate this work.

The first section of the paper describes the key issues in interference as the underlying motivation behind social rules. The next section puts forth the notion of prototypical social states which effectively prune the space of social interactions. The following part of the paper postulates the key sources of reinforcement in social learning. The reminder of the paper describes the robot experiments used to test the proposed theory as applied to learning social rules. The last part of the paper summarizes current conclusions and outlines continued experimental work.

2 Interference and Conflict

Although not universally accepted, a popular theory postulates that social interactions are evolutionarily beneficial because social structure minimizes interference and maximizes benefit of the group as a whole, presumably in terms of reproductive success.

Interference, then, is any influence that opposes or blocks the agents' goal–driven behavior. In societies consisting of agents with similar goals, interference manifests itself as competition for shared resources. In diverse societies, where agents' goals differ, more complex conflicts can persist between agents, including undoing of each other's work, deadlocks, and oscillations.

Two functionally distinct types of interference are relevant to this work: interference caused by multiplicity, which we will call *resource competition*, and interference caused by goal–related conflict, which we will call *goal competition*. Resource competition includes any interference resulting from multiple agents competing for common resources, such as space, food, and information. This type of interference causes the decline in performance in multi–agents systems as more agents are added. It is also the primary impetus for social rules.

While resource competition is simply caused by physical coexistence, and can thus exist in any multi–agent system, goal competition only applies to systems with heterogeneous agents. Functionally different agents can create lasting interference over conflicting goals. An agent can undo the work of another not out of immediate need for the resources but due to some higher–level goal such as, for instance, establishing dominance.

Goal competition is studied primarily by the Distributed AI community (Gasser & N. Huhns 1989). The approaches usually involve predicting other agents' goals and intentions, thus requiring agents to maintain models of each other's internal state (e.g. Huber & Durfee (1993), Miceli & Cesta (1993)). Such prediction abilities require computational resources that do not scale well with increased group sizes. In contrast, our work deals with homogeneous societies in which social rules are shared by all individuals. Consequently, only resource competition, and social rules aimed at minimizing it, are relevant and will be addressed.

2.1 Individuals vs. Groups

Social rules that minimize interference among agents attempt to direct behavior away from individual greediness and toward global efficiency.[1] Greedy individualist strategies perform poorly in group situations in which resource competition is inevitable and not correctly managed, since resource competition grows with the size of the group.[2] Not all group dynamics fall into this category. For example, some tasks allow for efficient greedy strategies in which agents specialize by task division, as demonstrated by Deneubourg, Goss, Pasteels, Fresneau & Lachaud (1987). In contrast, this work focuses on tasks and solutions in which the agents cannot specialize, and instead must find means of optimizing their activity within the same task by developing social rules.

In such situations, agents must give up individual optimality in favor of collective efficiency. At least in theory, it is in the interest of each of the individuals to obey social rules, since on the average their individual efficiency will be improved as well. However, since the connection between individual and collective benefit is not always direct, the problem of learning social rules is a difficult one.

Outside game theory, this problem has been addressed in the filed of Artificial Life at the genetic learning level. Genetic algorithm allow generations of agents with differ-

[1] Global efficiency can, in a cultural context, be elevated to "the common good."

[2] The rate of growth is determined by the properties of the particular system.

ent social rules to be tried out, mutated, and recombined, in order to find the best fit (for example, see MacLennan (1990) for a strategy for evolving communication, and Chris Langton (1989) for related work).

The work presented in this paper is fundamentally different because it addresses the problem learning of social rules by each of the individuals during their lifetime and within the context of a task, in this case foraging. It is aimed at loosely modeling higher animals rather than insects that are preprogrammed with the appropriate social strategies. Consequently, it cannot directly assume social rules and build them in, but rather it must deal with tradeoffs in individual versus global efficiency and incorporate those appropriately into the on–line real–time learning strategies.

3 Prototypical States in Social Interaction

Animal interactions are largely ritualized and drawn from a small social repertoire. In theory, the number of possible social situations among multiple agents is prohibitively large, especially since it grows with the size of the group. In practice, however, the rules of social behavior are quite independent of exact group size. Specifically, only a few classes of group sizes and prototypical relations among agents are relevant for any given type of interaction.

The canonical form of social relation is the *dominance hierarchy* or *pecking order*, ubiquitous in animal societies, from hermit crabs and chickens to primates and people (Chase 1982, Chase & Rohwer 1987). Besides mating, the majority of animal social interaction focuses on establishing and maintaining pecking orders (Cheney & Seyfarth 1990). Although no directly derivable evolutionary benefit has been proven, hierarchies certainly serve to ritualize and thus simplify social interactions.

While dominance hierarchies are a prevalent social structure and simplify social interaction, we are interested in learning social rules that are not directly embedded into a dominance structure, but instead need to be derived by the agent based on its interactions with others and its effectiveness at its tasks over time.

Besides by dominance hierarchies, animal social interactions can be classified into one-on-one, one-to-few, and one-to-many[3], with the first two being by far the most prevalent. The work described in this paper will use the group size classification to prune the state space of social interactions, and focus on learning social rules that are applied in one-on-one and one-to-few interaction classes. In particular, we will work on yielding rules for one-on-one motion conflicts, and communication rules for one-to-few interaction. We will postulate that learning social rules for such interactions requires specific types of social reinforcement, as described in the next section.

4 Social Reinforcement

Social settings offer a plethora of information useful for learning. The relevance of conspecifics' behavior is recognized in biology by the innate propensity for social learning. For instance, young animals copy the behavior of an adult automatically, often without any direct reward (Gould 1982). Furthermore, behavior of peers is also copied, especially if it leads to a reward. This type of social behavior results in the development of culture in the form of behavior patterns that are passed on through generations. In a well known example, monkeys repeatedly discovered washing potatoes and separating dirt from seeds (McFarland 1985).

Observed behavior of conspecifics can serve as negative as well as positive reinforcement. For example, animals quickly learn not to eat food that has had a poisonous effect on others, or to avoid people if others have been hunted by them (Davis 1973, Gould 1982, McFarland 1987). Vicarious punishment and vicarious reward are both effective, although less so then their direct counterparts. In a sense, vicarious learning is a means of distributing trials over multiple agents so that one agent need not perform then all. As long as the experience of an agent is "visible", it can serve as a source of vicarious learning trials for others.

Evidence from ethology guides us to derive three forms of reinforcement involved in social learning. The first type is the individual perception of progress relative to the current goal. This type of reinforcement is inherent in any learning task, but its availability varies depending on the specific agent and environment. In most tasks, the agent can maintain some measure of progress[4] that is critical for efficient learning.

The second type of reinforcement comes from observing the behavior of conspecifics. Similar behavior in a peer is considered to be a source of positive reinforcement. Although this in itself does not constitute a reliable reinforcement signal, coupled with a direct estimate of progress toward the agent's particular goal, it can provide useful feedback for the agent.

Finally, the third type of reinforcement is that received by conspecifics. Interestingly, using this type of information does not require the agent to model another agent's internal state. Since the agents belong to a homogeneous society in which they obey consistent social rules, any reward or punishment received by a conspecific would have been received by the agent itself in a similar situation. Consequently, vicarious reinforcement is a useful learning signal.

[3] Where "many" is bounded by the sensing and communication range.

[4] Most tasks do not resemble maze learning where the only reward is found at the goal.

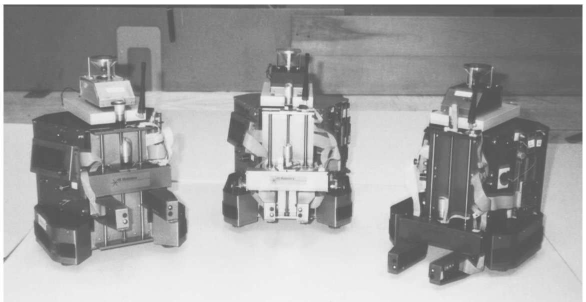

Figure 1: Each robot consists of a differentially steerable wheeled base, and a gripper that can grasp and lift objects. The robots' sensory capabilities include piezo–electric bump and gripper sensors, infra–red sensors for collision detection, proprioceptive sensors for maintaining drive and gripper–motor current, voltage, and position, and a radio transmitter for communication and absolute positioning.

We postulate that all three of the above described forms of reinforcement are necessary for learning social rules. Individual reinforcement is not sufficient because it, by definition, maximizes individual benefit. Since the kind of rules we are interested in learning do not immediately and directly benefit the individual and, in some cases, even have a delaying effect on its individual reinforcement, then cannot be learned with such reinforcement alone.

Observing other agents' behavior helps the learning because it gives the impetus to the agent to try behaviors that do not immediately benefit it. Repeating the behavior of others enforces multiple trials of an otherwise rare social behavior, so more stable reinforcement is received. Although observing other agents' behavior clearly encourages exploration in the group and thus speeds up learning, it still cannot induce individual agents to behave in ways that will reduce their reinforcement.

How can a society develop social rules based on individual learning, i.e. without some centrally imposed arbiter? It can be accomplished if the agents are able to estimate other agents' reinforcement and their individual reinforcement is positively correlated with their conspecifics. In short, what is good for one, is good for another, at least indirectly. This simple model allows for learning a variety of powerful social rules which minimize interference and maximize group benefit.

To test this hypothesis, we have designed a collection of experiments on situated agents (mobile robots) learning to yield and communicate, in order to become more globally efficient at foraging, their high–level task. The next section describes the experimental environment.

5 The Robots

The experimental environment to be described is designed for performing a variety of group behavior experiments. It has been used for verifying work on basic behaviors as well as for demonstrating an algorithm for accelerated situated learning (Matarić 1994). It allows for implementing various interactions between robots capable of communicating with each other, and sensing and manipulating objects.

The experiments are conducted on a collection of up to four R2 mobile robots. The robots are fully autonomous with on–board power and sensing. Each robot consists of a differentially steerable wheeled base and a gripper that can grasp and lift objects. The robots' sensory capabilities include a piezo–electric bump sensor strip around the base for detecting collisions, another strip inside each finger of the gripper for detecting the grasping force, and a set of infra–red (IR) sensors: two on the inside of the fingers for detecting graspable objects, and ten more for obstacle avoidance: a set of eight sensors around the top of the robot, and one more on each gripper tip (see Figure 1).

In addition to the described external sensors, the robots are also equipped with proprioceptive sensors supplying battery voltage and current information, sensors for drive motor current, and shaft encoders in the gripper motors for maintaining position and height information.

Finally, the robots are equipped with radio transceivers, used for determining absolute position and for inter–robot communication. Position information is obtained by triangulating the distance computed from synchronized ultrasound pulses from two fixed base stations. Inter–robot communication consists of locally broadcasting 6-byte messages at the rate of 1 Hz. The radios are particularly useful for transmitting any information that could not be reasonably sensed with the available on–board sensors, such as the external state of other robots (i.e. holding "food", finding home, etc.), which is required for social learning.

The robots are programmed in the Behavior Language, a parallel programming language based on the Subsumption Architecture (Brooks 1990, Brooks 1987). Their control systems are collections of parallel, concurrently active behaviors.

6 The Learning Environment

In order to make the learning task realistic, we tested the learning algorithms on robots situated in a foraging task, having a common higher–level goal of collecting "food" and taking it home during the "day" and sleeping at home at "night." Thus, the robots' world is a confined area which contains a fixed home region shared by all robots, and scaled so as to accommodate them, i.e. large enough for all of them to "park" for the night, with some

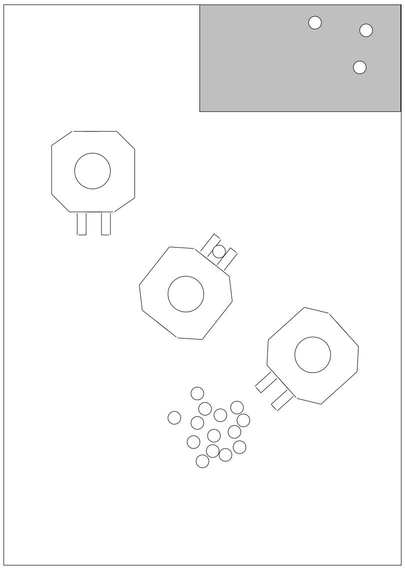

Figure 2: The experimental area in which the learning task was conducted. The home region is shaded. Pucks are clustered. The relative sizes of the environment and the robots are shown to scale.

reasonable amount of maneuvering (figure 2).

Food is clustered in a single location in the workspace. The clustered nature of the food certainly has an effect on what the optimal set of strategies will be for the collection task. We chose the single cluster arrangement for two reasons. First, clustering food gives incentive for the agents to cooperate and exchange information. If food were randomly scattered around the environment, there would be no point to communicating and social rules could be reduced to avoidance. Second, clustering food reduces the overall duration of each experimental run because the agents need to spend much less time looking for it. The complexity of the learning task is not diminished, but finite resources such as battery power are conserved by designing the experiment so as to maximize the relevance of the robots' interactions.

For the same reason, the chosen workspace is small enough to result in frequent interaction and interference between the robots in order to make learning possible.

The experimental environment is meant to loosely resemble a society that spends its days foraging (or hunting and gathering) by making repeated trips to a resource–rich area, getting the food, and taking it home. At fixed periods (meant to resemble night time), the society gathers at home and rests. Foraging activities resume at the beginning of each day.

7 The Learning Task

The task consisted of learning two social rules: 1) yielding and 2) communicating puck location. Sincy the rules are not symmetric, they translate into four behaviors, two per social situation. Yielding consists of learning when to on one side give way and on the other keep going. Sharing information consists of learning when to broadcast position information when when to receive and store it.

The social rules are expressed within the robots' natural habitat and the context of their usual routines, in this case foraging. The foraging context was chosen because our previous work (Matarić 1992, Matarić 1993) provided the basic behavior repertoire to which social rules could easily be added. The built–in foraging behavior consists of a finite state controller which, in response to mutually-exclusive conditions consisting of internal state, activates appropriate basic behaviors from the robots' repertoire. The conditions include externally–triggered events, such as getting close to another robot or finding a puck, and internally–generated events, such as the the onset of night time, indicated by the internal clock. Basic behaviors include avoidance, dispersion, searching for pucks, picking up pucks, homing, and sleeping. Our earlier work has shown how foraging itself can be learned (Matarić 1994). This work demonstrates how it can be improved with social rules.

It is important to note that the alternative to behaving socially is not behaving randomly, but behaving greedily.[5] Thus, this learning task is realistic and difficult, since each agent is biased toward its built–in instinctive greedy and individualistic behavior which guarantees survival.[6]

In order to learn social rules, the behavior repertoire must include the appropriate social behaviors: yielding, proceeding, broadcasting, listening & storing, and observing & remembering, so that these behaviors would be tried and eventually have their conditions learned. Most of the above active behaviors did not require adding new abilities to the robots, since they already contained the necessary actions for each: yielding consists of stopping and waiting, and proceeding consists of going on with the current behavior. Non-active behavior requiring observation and storage were added exclusively for the social learning algorithms, but are based on the existing mechanism for sending and receiving messages already used for communicating positions of neighbors. The job of the learning algorithm was to correlate appropriate conditions for each of these behaviors in order to optimize the higher–level behavior, i.e. to maximize received reinforcement.

If this learning task could be set up as a simple search in the space of possible strategies, with highest reward for the strategies that result in highest efficiency of foraging, then it could be simply solved using game–

[5] Greediness is often the basis of antisocial behavior.

[6] No interpretation of basic human nature is implied or intended. Tendency toward friendliness can also be built–in but is not the goal of this experiment.

theoretic strategies that derived tit–for–tat related solutions. However, in the described domain, the agents cannot simply "search" the condition–behavior space because they cannot directly and completely control events in the world. Instead, the world is stochastic, noisy and uncertain, and in particular it is made more complex by the presence of other learning agents. However, it is exactly their presence and the ways in which they interact that constitute relevant conditions for learning social rules.

We postulate the following three types of social knowledge and related reinforcement as necessary for learning the proposed social rules:

1. direct reinforcement,
2. observation of the behavior of other agents,
3. observation of reinforcement received by other agents.

The next section describes how this reinforcement was formally implemented.

8 Implementing Social Reinforcement

The learning algorithm produces and maintains a total order on the appropriateness of social behaviors associated with every condition, expressed as $A(c,b)$. As described earlier, the matrix of $A(c,b)$ values is not prohibitively large, since the number of relevant social conditions and behaviors is relatively small. The values in the correlation matrix fluctuate over time based on the received reinforcement, but the learning algorithm is expected to converge on a stable ordering. At each point in time the value of $A(c,b)$ is the sum of all past reinforcement R:

$$A(c,b) = \sum_{t=1}^{T} R(t)$$

The implementation of direct reinforcement is straightforward. A progress monitoring behavior constantly compares the agent's current state with its immediate goal. Whenever it detects progress (whether in terms of reaching a subgoal, as in the case of finding food, or in terms of diminishing the distance toward the goal, as in the case of going home), it gives a small reward to the most recently active social behavior.[7] Analogously, whenever a regress from the goal is detected, a small punishment is delivered. Formally:

$$D(t) = \begin{cases} m & \text{if progress is made} \\ n & \text{if regress is made} \\ 0 & \text{otherwise} \end{cases} \qquad m > 0, \quad n < 0$$

[7]Note that only *social* behaviors are reinforced, since only social rules are being learned.

This algorithm is guaranteed to keep the learning system from settling in local minima, since the system continuously adapts the correlation values with each received reinforcement. The algorithm relies on estimating progress at least intermittently. If progress measurements are not available, and the reward function is an impulse at goal only, then the algorithm reduces to one–step temporal differencing (Sutton 1988), which has been proven to converge, however slowly. Given that most environments are not of this kind, this algorithm is bound to work faster. It has been successfully used to acquire the basic, greedy strategy for individual foraging (Matarić 1994).

The motivation for using progress estimators (also called internal critics) rather than a delayed reinforcement algorithm such as Q-learning comes from the non-Markovian uncertainty and inconsistency properties of the group situated agent domain. The reasoning is described in detail in Matarić (1994) but in short, progress estimators provide unbiased, principled reinforcement which is not so delayed as to become useless in a dynamic environment.

Observational reinforcement, i.e. reinforcement for repeating another agent's behavior is delivered in a similar form:

$$O(t) = \begin{cases} o & \text{if repeated observed behavior} \\ 0 & \text{otherwise} \end{cases} \qquad o > 0$$

An agent receives positive reinforcement if it repeats the behavior it has most recently observed next time it finds itself in the same condition, unless it has already recently performed that behavior. $O(t)$ has a temporal component. It expires after a fixed time so that the agent, in effect, forgets what it last saw if it has not seen it recently. This feature eliminates some cyclic behavior patterns between multiple learning agents observing each other.

Finally, vicarious reinforcement, i.e. reinforcement received based on that received by other agents, is delivered in the following form:

$$V(t) = \begin{cases} v & \text{if vicarious positive reinforcement} \\ w & \text{if vicarious negative reinforcement} \\ 0 & \text{otherwise} \end{cases}$$

$$v > 0, \quad w < 0$$

Vicarious reinforcement delivers a form of "shared" reinforcement to all agents involved in a local social interaction. By spreading individual reward or punishment over multiple agents, it serves to extend individual benefit beyond a single agent. As a consequence, the amount of reward received for social behaviors over time outweighs that received for greedy strategies.

The complete reinforcement function then, is the sum of the subset of social reinforcement being used in the given learning experiment. Specific examples are given in the next section.

9 Learning Experiments

In a typical learning experiment all of the agents are endowed with identical basic behaviors and the social learning reinforcement function. They are started out from home at a beginning of a "day" and go about their foraging task as usual. The social learning algorithm is activated whenever an agent finds itself:

1. near a large amount of food away from home,
2. receiving an agent's message,
3. within observing range of another stopped agent,
4. within observing range of another moving agent,
5. within interference range of another stopped agent,
6. within interference range of another moving agent.

These conditions are specified by the designer in order to speed up the learning. They could be learned using one of the available statistical methods for state generalization (for example, Chapman & Kaelbling (1991) and Mahadevan & Connell (1991)) but the process could take almost as long as the social learning and is likely to suffer from sensory errors.

The first condition enables learning to communicate about sharable resources such as food. The last two conditions are based on two distance thresholds established a priori: $\delta_{observe}$ and $\delta_{interfere}$. In the basic foraging behavior, the presence of another agent within $\delta_{interfere}$ triggers avoidance. In the social learning algorithm, a social behaviors is attempted as an alternative.

Whenever an event occurs that makes one of the above conditions true, the current behavior the agent is executed is terminated and appropriate social reinforcement is delivered. Then, a new behavior is selected using the following strategy:

1. choose a recently observed behavior, if appropriate,
2. choose an untried social behavior,
3. choose the best behavior,
4. choose a random behavior (5% probability)

Given their confined physical work area, the agents interact enough to generate a plethora of events enabling social conditions. Consequently, we can observe learning in real time over the period of minutes.

10 Preliminary Results

As mentioned previously, the basic learning algorithm has been tested and shown to work both theoretically and practically, and the results are described in detail in Matarić (1994). Within an average fifteen minute trial run the robots were able to learn foraging by learning to select appropriate individual behaviors for each state. The algorithm, however, did not allow for learning social rules, which is why social reinforcement was introduced.

Condition	Behavior
finding food	broadcast
receiving a message	store location
near a stopped agent	proceed
near a moving agent	store behavior
too near a stopped agent	proceed
too near a moving agent	yield

Figure 3: The condition–behavior pairings of the desired social policy.

We tested the following reinforcement functions:

1. $A_d(s,b,T) = \sum_{t=1}^{T} D(t)$
2. $A_{do}(s,b,T) = \sum_{t=1}^{T} (\alpha D(t) + \beta O(t))$
3. $A_{dov}(s,b,T) = \sum_{t=1}^{T} (\alpha D(t) + \beta O(t) + \gamma V(t))$

Weighting factors in learning rules are usually ad hoc. Our approach uses a simple scheme in which direct progress is weighted the highest, while observation–induced experience and vicarious reinforcement are weighted less. In the second set of experiments listed above we used:

$$\alpha + \beta = 1 \;\; and \;\; \alpha > \beta$$

In the third set we used:

$$\alpha + \beta + \gamma = 1 \;\; and \;\; \alpha > \beta \;\; and \;\; \beta = \gamma > \gamma$$

The relative effectiveness of the three reinforcement types was evaluated based on the portion of the social rule policy the algorithms learned. Figure 3 shows the condition–behavior pairings of the desired social policy.

The evaluation metrics used to compare the results of the different reinforcement were:

1. the percentage of the optimal social strategy the agents managed to learn,
2. the relative amount of time required to learn it.

Relative rather than absolute time is used because the duration of any learning run varies depending on the arrangement of the agents and the temporal distribution of their interactions. Consequently, exact time to convergence is not a valid evaluation metric in an event–driven situated environment of the kind we are using. Relative average time, on the other hand, gives us a measure of comparative effectiveness of different algorithms.

As a control experiment, we tested the performance of a pre-programmed foraging behavior that contained the

Reinforcement	Performance
$A_d(s, b, T)$	does not converge
$A_{do}(s, b, T)$	does not converge
$A_{dov}(s, b, T)$	converges

Figure 4: Comparative performance of different reinforcement functions.

desired social rules, and compared it to the base case foraging. Not surprisingly, groups using social rules always outperformed groups with only greedy individual strategies. Thus, we were convinced that incentive for learning social rules did exist, and it was then a matter of finding out which reinforcement strategy makes it learnable in the given environment.

Figure 4 shows preliminary convergence results. Convergence is defined as learning the complete social policy. The results shown above are averaged over multiple trials and are qualitative because insufficient trials were run to perform accurate statistical analysis. However, results from the performed trials were consistent in that the first two strategies never converged, and learned only a very small part of the policy. The third strategy converted in all sufficiently long trials, i.e. some trials required up to 20 minutes to converge. Trial duration was not an issue in case of the other two reinforcement strategies, as they did not improve over time and uniformly failed to converge regardless of experiment duration.

Duration of learning was a direct effect of using physical hardware to run the learning experiments, since the domain was ridden with unavoidable intermittent error and noise. For instance, agents did not always behave consistently due to their inability to accurately sense external state. Such unavoidable (and undetectable) errors generated "unintentional deserters" in that robots experiencing sensor error (and in some cases radio transmission delay in communication) occasionally failed to behave socially. We found that these effects did not disable the learning algorithm, although they did slow it down, as described above. The basic learning algorithm is designed to adapt throughout the lifetime of an agent so learning does not suffer long–term consequences from intermittent errors.

In terms of speed of learning, the social condition–behavior pairs consistently ranked in the following increasing order of difficulty, as shown in figure 10. This ordering directly reflects the immediacy of the reward associated with each social behavior. The rules that produce the most immediate reward were learned the fastest. The social rule of sharing information about food was by far the hardest to learn, as the benefit to the agent had the least direct payoff. Consequently, it required multiple instances of observational reinforcement, which instigated exploration of the behavior. If after attempting it, the agent could not received immediate vicarious reinforcement since other agents, who received the message, had not always learned to store and use the information, thus receiving positive reinforcement and passing it on to the agent who originally broadcast the information. This effect could be eliminated if agents were learning from experts, but we were interested in having social rules emerge in a homogeneous group.

Condition	Behavior
near a stopped agent	proceed
too near a stopped agent	proceed
receiving a message	store location
near a moving agent	store behavior
too near a moving agent	yield
finding food	broadcast

Figure 5: The Relative difficulty of each condition–action pair, in increasing order.

The proposed observational and vicarious reinforcement strategies, coupled with direct reinforcement, were effective in enabling the learning of social rules, but their effectiveness varied depending on the rule being learned. This is not unexpected, since some rules are certainly harder than others. However, the difficult of certain learning problems, such as particularly altruistic social rules, indicates that those are best learned genetically. Data from biology seems to support this intuition, since animals do not appear to learn altruism toward their kin but instead are endowed with it (McFarland 1985).

All of the presented results are preliminary, and are only a glimpse at the wide variety of social rules that can be learned, and forms of social reinforcement that are worth exploring. In order to properly evaluate our theories, we are currently implementing a battery of tests which will produce an amount of learning data that allows us to draw conclusions that are statistically significant.

11 Summary and Continuing Work

This work has focused on learning social rules in situated multi–agent domains. We have studied the difficulty of learning behaviors which do not have direct benefit to the agent and therefore are in contradiction with its basic, greedy, survival instincts. We postulated three types of reinforcement that are useful and possibly necessary for learning such social rules. We then tested this postulate by applying an already effective situated learning algorithm to the social learning problem by adding the proposed types of reinforcement to it. We demonstrated the algorithms by implementing them on a group of four autonomous mobile robots capable of communication and cooperation and given the task of learning yielding and sharing information.

We are interested in expanding this work in a number of directions. In particular, it would be interesting to consider variations on the learning experiments, such as a gathering task with multiple food and home regions, in order to study what kinds of specializations emerge between agents and how those affect the resulting social rules. We would also like to test the presented social reinforcement strategies on quite different types of tasks in order to see how general they are. Another area we are interested in exploring is learning to distinguish what aspects of the situation (state) are relevant, in the context of learning relevant group sizes and dominance hierarchies. If this turns out to be difficult to learn it will give us an idea of what types of biases may be genetically programmed.

The work we have presented is very much in progress. The data for comparison and evaluation of the different learning strategies is still pending and is being gathered gradually. A full statistical analysis of a complete set of experiments is in order and is in preparation. We believe that the presented preliminary results give us new insight into how difficult certain types of social learning may be and how we may go about synthesizing it on artificial agents in order to study biological mechanisms of interaction and social learning.

Acknowledgements

The author wishes to thank Simon Goss and an anonymous reviewer for detailed helpful comments on an earlier draft of this paper.

The research reported here was done at the MIT Artificial Intelligence Laboratory. Support for this research was provided in part by the Jet Propulsion Laboratory contract 959333 and in part by the Advanced Research Projects Agency under Office of Naval Research grant N00014–91–J–4038.

References

Axelrod, R. (1984), *The Evolution of Cooperation*, Basic Books, New York.

Brooks, R. A. (1987), A Hardware Retargetable Distributed Layered Architecture for Mobile Robot Control, *in* 'IEEE International Conference on Robotics and Automation', Raleigh, NC, pp. 106–110.

Brooks, R. A. (1990), The Behavior Language; User's Guide, Technical Report AIM-1127, MIT Artificial Intelligence Lab.

Chapman, D. & Kaelbling, L. P. (1991), Input Generalization in Delayed Reinforcement Learning: An Algorithm and Performance Comparisons, *in* 'Proceedings, IJCAI-91', Sydney, Australia.

Chase, I. D. (1982), 'Dynamics of Hierarchy Formation: The Sequential Development of Dominance Relationships', *Behaviour* **80**, 218–240.

Chase, I. D. & Rohwer, S. (1987), 'Two Methods for Quantifying the Development of Dominance Heirarchies in Large Groups with Application to Harris' Sparrows', *Animal Behavior*.

Cheney, D. L. & Seyfarth, R. M. (1990), *How Monkeys See the World*, The University of Chicago Press, Chicago.

Chris Langton, e. (1989), *Artificial Life*, Addison–Wesley.

Davis, J. M. (1973), Imitation: A Review and Critique, *in* Bateson & Klopfer, eds, 'Perspectives in Ethology', Vol. 1, Plenum Press.

Deneubourg, J. L., Goss, S., Pasteels, J. M., Fresneau, D. & Lachaud, J. P. (1987), 'Self-Organization Mechanisms in Ant Societies, II: Learning in Foraging and Division of Labor', *From Individual to Collective Behavior in Social Insects* **54**, 177–196.

Durfee, E. H., Lee, J. & Gmytrasiewicz, P. J. (1993), Overeager Reciprocal Rationality and Mixed Strategy Equilibria, *in* 'Proceedings, AAAI–93', Washington, DC, pp. 225–230.

Fiority, G. & Scotto, P. (1992), 'Observational Learning in Octopus vulgaris', *SCI* **256**, 545–547.

Gasser, L. & N. Huhns, e. (1989), *Distributed Artificial Intelligence*, Pitman, London.

Gould, J. L. (1982), *Ethology; The Mechanisms and Evolution of Behavior*, W. W. Norton & Co., New York.

Huber, M. J. & Durfee, E. H. (1993), Observational Uncertainty in Plan Recognition Among Interacting Robots, *in* 'Proceedings, IJCAI-93 Workshop on Dynamically Interacting Robots', Chambery, France, pp. 68–75.

Kraus, S. (1993), Agents Contracting Tasks in Non-Collaborative Environments, *in* 'Proceedings, AAAI-93', Washington, DC, pp. 243–248.

MacLennan, B. J. (1990), Evolution of Communication in a Population of Simple Machines, Technical Report Computer Science Department Technical Report CS–90–104, University of Tennessee.

Maes, P. & Brooks, R. A. (1990), Learning to Coordinate Behaviors, *in* 'Proceedings, AAAI-91', Boston, MA, pp. 796–802.

Mahadevan, S. & Connell, J. (1991), Automatic Programming of Behavior-based Robots using Reinforcement Learning, *in* 'Proceedings, AAAI-91', Pittsburgh, PA, pp. 8–14.

Matarić, M. J. (1992), Designing Emergent Behaviors: From Local Interactions to Collective Intelligence, *in* 'From Animals to Animats: International Conference on Simulation of Adaptive Behavior'.

Matarić, M. J. (1993), Kin Recognition, Similarity, and Group Behavior, *in* 'Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society', Boulder, Colorado, pp. 705–710.

Matarić, M. J. (1994), Reward Functions for Accelerated Learning, *in* 'Proceedings, Eleventh International Conference on Machine Learning (ML-94)'.

McFarland, D. (1985), *Animal Behavior*, Benjamin Cummings.

McFarland, D. (1987), The Oxford Companion to Animal Behavior, *in* 'Oxford, University Press'.

Miceli, M. & Cesta, A. (1993), Strategic Social Planning: Looking for Willingness in Multi-Agent Domains, *in* 'Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society', Boulder, Colorado, pp. 741–746.

Mitchell, R. W. (1987), A Comparative-Developmental Approach to Understanding Imitation, *in* Bateson & Klopfer, eds, 'Perspectives in Ethology', Vol. 7, Plenum Press.

Moore, B. R. (1992), 'Avian Movement Imitation and a new Form of Mimicry: Tracing the Evolution of a Complex Form of Learning', *Behavior* **122**, 614–623.

Sutton, R. (1988), 'Learning to Predict by Method of Temporal Differences', *The Journal of Machine Learning* **3**(1), 9–44.

Tomasello, M., Kruger, A. C. & Rather, H. H. (1992), 'Cultural Learning', *to appear in The Journal of Brain and Behavior Sciences.*