Evolution of cooperation within a behavior-based perspective: confronting nature and animats*

Samuel Delepoulle[1,2] delepoulle@univ-lille3.fr,
Philippe Preux[2] preux@lil.univ-littoral.fr,
Jean-Claude Darcheville[1] darcheville@univ-lille3.fr

[1] Unité de Recherche sur l'Evolution du Comportement et des Apprentissages, Université de Lille 3, B.P. 149, 59653 Villeneuve d'Ascq Cedex, France.
[2] Laboratoire d'Informatique du Littoral, B.P. 719, 62228 Calais, France.

Abstract. We study the evolution of social behaviors within a behavioral framework. To this end, we define a "minimal social situation" that is experimented with both humans and simulations based on reinforcement learning algorithms. We analyse the dynamics of behaviors in this situation by way of operant conditioning. We show that the best reinforcement algorithm, based on Staddon-Zhang's equations, has a performance and a variety of behaviors that comes close to that of humans, and clearly outperforms the well-known Q-learning. Though we use here a rather simple, yet rich, situation, we argue that operant conditioning deserves much study in the realm of artificial life, being too often misunderstood, and confused with classical conditioning.

1 Motivations: Artificial Life and the experimental analysis of behavior

As stated at its origins, artificial life deals with the study of life as it exists, and life as it might exist [18]. This paper deals with the first part of this project, the study of life as it exists. Our endeavor concerns the study of the dynamics of behavior relying on simple, though sound, nature grounded, assumptions. These assumptions are drawn from a selectionist approach of behaviors, compatible with, and complementary to, the selectionist approach of the evolution of living species. Basically, relying on Skinner's work, the selectionist approach of behaviors (or radical behaviorism, operant or instrumental conditioning [24, 23, 26]), states that a behavior is more likely to be re-executed, and spread in the population, if it is followed by positive consequences. Conversely, a behavior is likely to disappear if it is followed by negative consequences. Furthermore, behaviors being generally never redone twice exactly identically, variation naturally occurs in behaviors, hence a mutation-like mechanism is naturally included. This has much to share with the idea of genes being retained and spreading in a population if they are well adapted [25]. Operant learning is well known to be

* This research is supported by "Conseil Regional Nord-Pas de Calais" (contract n 97 53 0283)

adequate to explain complex behaviors of animals, and to let animals acquire complex behavior [26]. These behaviors are indeed much more complex than those currently exhibited by top of the art robots [5]. In [7], we have shown that a task consisting in sharing a task between agents can be solved quasi-optimally by a set of agents which behavior evolution is selectionist. In this paper, we focus on the emergence of cooperation among a set of agents and explain it by mean of a selection of behaviors due to the context in which they are emitted, or "contingencies of reinforcement". Our approach is original is many respects. First, we try to keep a pure behavior-based approach, relying on the work of the experimental analysis of behavior. Second, we work in the same time on living beings, mostly human beings, and computer simulations. Computer simulations compel us to define very precisely the principles of selection of behaviors and give us the opportunity to simulate them very strictly, showing their validity, and showing their weaknesses. This research might lead to new ideas on the issue of making cooperation between adaptive agents possible. In the world of agents, cooperation is considered as a main issue, and as a problem to solve in itself. Stemming from the analysis of behavior, we wish to put forward the point of view that cooperation can also be considered as a "natural" adaptation of agent's behavior to their environmental contingencies, as far as the agents are indeed able to adapt themselves. Furthermore, we discuss below the fact that cooperation might not be regarded as a so-well formulated problem.

2 Evolution of cooperation

What makes the evolution of cooperation possible between two, or more, living beings? In natural situations, one living being frequently behaves in such a way that produces favorable consequences for other living beings. For the individual who emits such a behavior, the immediate payoff can be inexistent, or negative. Hence, the individual does not directly benefit from his, or her, own behaviors. How can such behaviors appear? How can they be retained? These are serious problems arising if we adopt a selectionist way of thinking, such as within the theory of natural selection, or the selection of behaviors by way of their consequences [25].

In his famous work [2], Axelrod uses the problem of the Iterated Prisoner's Dilemma (IPD) to formulate the problem of cooperation in the framework of game theory. He demonstrates the importance of the repetition of cooperation situations. Dawkins [6] studies the role of genes in cooperative situations. Biologists works show the importance of social behaviors [11, 34] Many other attempts to explain cooperative behaviors have been made [16, 12]. However, none of them have yet tackled the question of the evolution of cooperation within a strict behavioral perspective. In this paper, we study the conditions within which cooperation can appear and get installed.

We propose to examine a rather simple situation in which cooperation can appear. We do not work with the IPD because we think IPD is inadequate for our purpose with several respects. In the IPD,

- agents act synchronously: at each time slot, they perform an action. However, synchronization has a far from neglectable effect on the dynamics of behaviors as is shown below, so that synchronicity should emerge rather than being a pre-requisite of the situation,
- agents do not have the choice of doing nothing for a while, they always have to do something,
- the consequence, positive, or negative, is correlated straightforwardly to the action they chose at the previous time slot,
- no behavior has, or seems to have, no consequence, either positive or negative,
- agents are perfectly aware[1] of the potential consequences of their behaviors.

Different points of crucial importance have been let apart with regards to more realistic situations. In nature:

- the consequence of a behavior may be delayed so that the relationship between a behavior and its consequence gets blurred and eventually hard to figure out,
- lots of behaviors are not rewarded: they are neutral with regards to their consequences, hence their selection. Due to the parallelism between the selection of behaviors and the selection of genes, we emphasize this point because a crucial importance of "neutral mutation" in the evolution of living species is suspected.

Dealing with the study of the evolution of cooperation in natural/real-world situations, we design an experimental situation relaxing IPD limitations and involving the previous points. According to a straightforward Skinner's heritage, cooperation naturally emerges, if it ever does, thanks to an exchange of reinforcements between the agents[2].

We have worked along a line that confrontates the dynamics of behavior of human beings with that of animats merely implementing reinforcement algorithms. We put a fundamental emphasis on this confrontation for different reasons. To be clear, we obviously do not have the ambition to simulate a human being (!): the algorithms that we use are much cruder. More seriously, for the problem of cooperation in a certain experimental layout, we aim to compare the dynamics of human behaviors with that of reinforcement algorithms. Then, relying on the fact that the algorithms are all selectionist, we suggest that human beings may behave in the same situation according to selectionist principles if both dynamics match. Furthermore, this would suggest that a social behavior can be the consequence of individual behaviors. The aim is also to assess the performance of reinforcement algorithms in environments in which we try to carefully keep the very essential features of the real world.Based on

[1] as far as a virtual agent can be aware of something...

[2] it is remarkable that cooperation emerges at a rather identical rate and pace whether the human subjects have been completely told the situation before the experiment, or not, the latter being the normal way the experiment has to be conducted.

insights taken from naturally evolved structures, these algorithms are seldomly put into environments that somehow mimics the environment they evolved in and to which they have adapted. We think that some "weird"or deceiving performance of these algorithms when used in artificial environments (related to combinatorial optimization for instance) is simply due to the fact that they are suited to the complexity of the environment in which they evolved and not to the artificial environments in which we sometimes use them.

In the following, we first develop on the idea of cooperation and we describe the experimental situation we have designed. Then, we briefly describe the reinforcement algorithms that were used, most of them being directly inspired by principles of the analysis of behavior. Based on simulations, we discuss their ability to evolve cooperation. Finally, we draw some conclusions and discuss various perspectives of this work.

3 A behavioral perspective on human cooperation

3.1 Description of a "minimal social situation"

As stated by Hake and Vuklich [10] within a behavioral framework, cooperation is defined by the fact that the reinforcement of two individuals must be *"at least in part dependent upon the responses of the other individual"*. In cooperative procedures, an individual can improve not only his own payoff but also the payoff of his, or her, party.

In order to understand the development of cooperation, we use a very simple situation, namely a "minimal social situation". In that situation, subjects (humans) can interact through a controlled device, a computer. Each response of a subject is made on a computer, each subject being located in a separate room, and being unaware of the presence of his, or her, party. Thus, they are not told that they are in a social situation. As a matter of fact, most subjects claim at the end of the experimentation that they were alone in front of a computer, without any interaction of any kind with anything (*e.g.* an animat), or anyone else.

The experimental layout is directly inspired by Sidowski's experiments [21, 22]. In this situation, subjects are invited to interact with a computer during thirty minutes. On the computer screen, a window includes two buttons and a counter. When subject A (resp. B) clicks on button 1, B's counter (resp. A's) is incremented (rewarding action, or R action). When A (resp. B) clicks on button 2, B's counter (resp. A's) is decremented (punishing action, or P action). Then, the behavior of a subject has no result on his, or her, own counter but on the counter of his, or her, party. It is noteworthy that a subject may choose to do nothing for a while, or even during the whole experiment. Furthermore, agent actions are not synchronous: they do not have to click at the same time and one click is immediately taken into account to provide its consequence to the other subject. This is a fundamental difference between the situation we use and other published work in the field.

3.2 Results

Thirteen couples of subjects underwent this experiment. During the first minutes of the experiment, clicks are equally shared on the two buttons but after a few minutes the rate of clicks on the button 1 increases. At the end of the experiment, subjects click about five times more on button 1. These results accord with other studies using similar procedures with humans [21, 22], and with animals [4].

3.3 Interpretation

According to the experimental analysis of behavior [24, 26] we can suggest how cooperative behaviors emerge. In this approach, behaviors are selected by their consequences [9]. The behavior of an organism varies; if these variations are profitable to this organism, they are likely to be selected.

In the experiment we made, if the action (R, or P) of both subjects are done synchronously, three combinations are possible.

- If both subjects press button 1, each one gains 1 point. In that case, clicking on button 1 is reinforced and they will continue clicking on button 1,
- If both subjects press button 2, each one loses 1 point. So, they will change their behavior and press button 1. Then, the subjects comes to the previous situation,
- If one subject presses button 1 while the other one presses button 2, the first subject loses 1 point while the second gains 1 point. The first will change his, or her, behavior and the second will not. Then, subjects are in the second situation, which again leads to the first.

B\A	clicks on button 1		clicks on button 2
clicks on button 1	+1 +1		+1 -1
		↖	↓
clicks on button 2	-1 +1	→	-1 -1

Table 1. This table gives the payoff for both subjects of their actions. The first payoff (±1) of the couple is that of A, the second is that of B. See the text for explanations of the arrows.

Table 1 displays the consequences of the subject's choices if they act synchronously. If their actions are synchronous, the spaces of states is made of 4 states. Arrows show the trajectory among the states. State (+1 +1) is an attractor for all initial states. Thus, cooperation is an attractor for the dynamics of behaviors in this experimental situation: if agents behaves synchronously, cooperation appears very quickly and remains. We observe this phenomenon experimentally: originally behaving asynchronously, cooperation appears in the group exactly at the very moment when responses become synchronized.

Each pair of subjects can be represented in a two dimensional space, the axis being the cooperative rate of the two subjects. Figures 1(a) and 1(c) represent the evolution of the cooperative rate of the 13 couples of subjects. In figure 1(a) representing the begining of the experiment, we notice that reponses form a two dimensional gaussian-like distribution at the center. That distribution may be the result of random responses of the two subjects. Figure 1(c) shows the result during the four last minutes of the experiment. The major part of the distribution is concentrated on a corner, this point represents the maximum cooperation rate for both subjects. Considering the results of each group, we notice that when a pair of subjects is entering in that state, they always stay in it. After a certain amount of time (very variable), most of groups switch to that state.

To finish the analysis of the situation, we notice that a mere stimulus-response architecture is not able to solve the minimal social situation even in this very restricted case where each agent can either reward or punish his, or her, party, and actions are synchronous. A response-stimulus architecture is required which opens up the road to operant conditioning.

4 Reinforcement learning algorithm facing cooperation

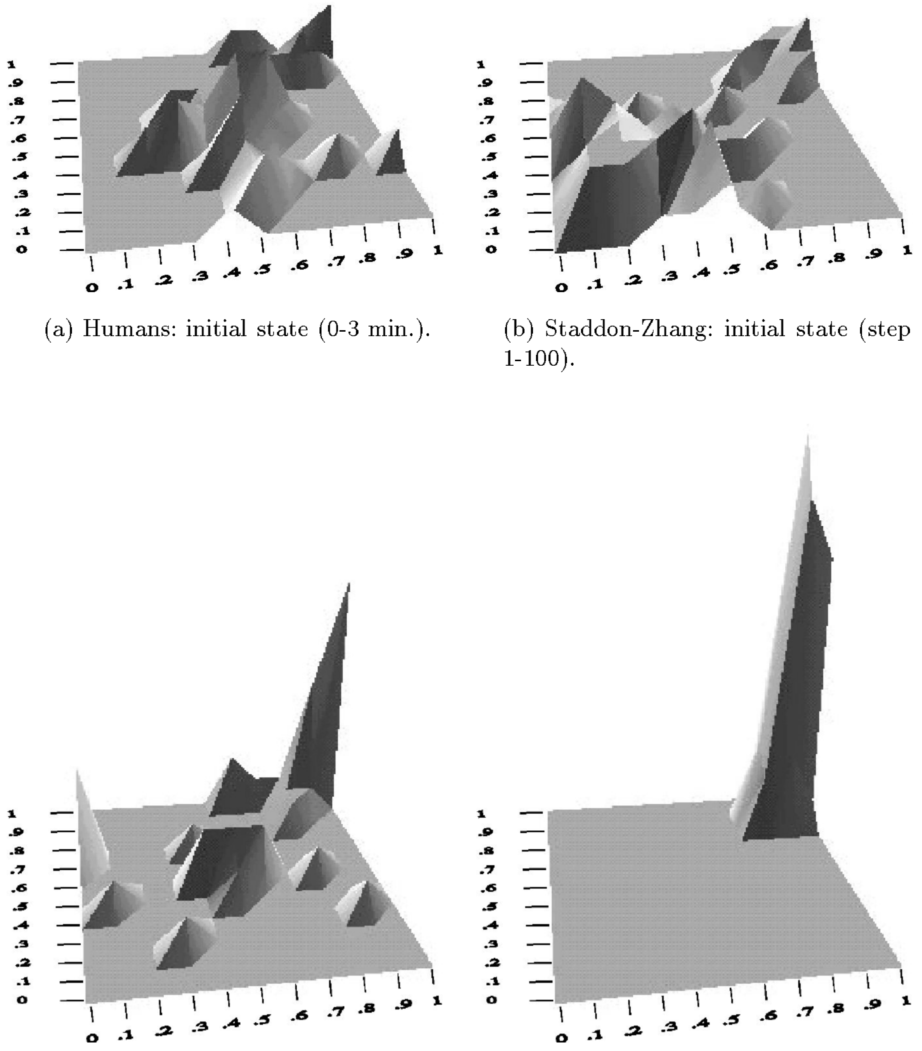
The interest of the experiment is not to show that human subjects can learn to behave in very simple social situations because we know that humans can learn much more complex social situations like imitation or verbal behavior. The interest is to show that a social situation can be explained by the knowledge of the individual law of behavior. Social organization can be the emergent result of individual behaviors. This analysis can also be supported by experimental work on social insects. It was shown that insects living in societies (ants for example) can perform complex tasks without the presence of an individualized central organizer [8, 30]. Many works have used this model to construct self-organised population of agents (see [19, 20] for instance).

The previous section has shown how the behavior of humans subjects can be explained using the principle of behavioral selection. So we propose to use agents based on a selectionist architecture. Reinforcement learning algorithms (RL) meet this requirement [17].

We have used five reinforcement-learning architectures, four of them being directly originated from the analysis of behavior while the fifth is the well-known Q-learning algorithm.

An agent can emit only three behaviors. It can reinforce its party (R), punish its party (P), or do nothing (N). We introduce the "do nothing" behavior to have an asynchronized situation at the beginning of the simulation. As we have shown in the experimental section, synchronization makes selection of cooperation easier because the consequence of a behavior is received immediately.

In the sequel of this section, we present the 5 reinforcement learning algorithms very briefly and give the experimental results in the next section.



(a) Humans: initial state (0-3 min.).

(b) Staddon-Zhang: initial state (step 1-100).

(c) Humans: final state (26-29 min.).

(d) Staddon-Zhang: final state (step 901-1000).

Fig. 1. Graphical representation of the cooperative rate of the couples of subjects (1^{st} column) and Staddon-Zhang simulation (2^{nd} column) at the beginning of an experiment (a), and at the end of the experiment (c). The abscissa is the rate of clicks on button 1 of subject, or agent, A, the ordinate is the rate of clicks on button 1 of subject, or agent, B. At the beginning of the experiment, the distribution is rather gaussian at the center, meaning that clicks are somehow equally shared between the two buttons. At the end of the experiment, there is a clear spike corresponding to the cooperation: both subjects, or agents, click only on button 1. Of course, some noise always occurs which explains the little blobs here and there.

4.1 The law of effect

At the end of the XIX$^{\text{th}}$ century, Thorndike [31,32] has studied the animal behavior and he has suggested a law to predict the evolution of behaviors with regards to their consequences (positive or negative). We formalize his "law of effect" as follows:

$$\text{if } C_i \neq 0, \text{let } s = \frac{C_i.\alpha}{|C_i|}, \text{then} \begin{cases} p_i = \frac{p_i+s}{1+s} \\ p_j = \frac{p_j}{1+s} \text{ for } j \neq i \end{cases}$$

where p_i is the probability of apparition of a behavior i, C_i, the consequence of behavior i and α is the learning rate, that is the relative weight of current stimuli with regards to behaviors emitted previously.

4.2 Hilgard and Bower's law

In Hilgard and Bower's law [13], very similar to the law of effect, named "linear reward–inaction algorithm", all non-reinforced actions are weakened. This algorithm always converges with a propability 1 on a particular action (but not always on the best action). This can be expressed by:

$$\text{if } C_i > 0 \begin{cases} p_i = p_i + \alpha(1 - p_i) \\ p_j = p_j - \alpha p_j \qquad \text{for } j \neq i \end{cases}$$

with p_i, C_i and α defined as in the law of effect.

4.3 Staddon-Zhang Model

In 1991, Staddon and Zhang [27] proposed a model in order to solve the *assignment-of-credit* problem without teacher (unsupervised learning). Staddon and Zhang show that this model accounts for qualitative properties of response selection. Their model accords itself not only with regular data in behavior analysis but also with "abnormal" behaviors (autoshaping, superstition and instinctive drift).

In this model, each behavior has a value V_i. All values V_i are in competition and a "winner-take-all" rule is used to select the behavior to emit at each time slot. The values V_i are updated at each time slot according to the equation:

$$V_i = \alpha V_i + \epsilon(1 - \alpha) + \beta V_i$$

where $0 < \alpha < 1$ (α is a kind of short term memory parameter), β, the reinforcement parameter should be positive for rewards and negative for punishments, and ϵ is a white noise.

4.4 Action-value method

The goal of this method is to estimate the mean consequence for each behavior and to choose the best one to emit in order to optimize the reward. Sutton [29] gives an iterative method to calculate this estimation. V_i is the estimation of mean consequence and N_i is the number of occurences of behavior i in the past. If behavior i is emitted, then

$$\begin{cases} V_{i+1} = \frac{1}{N_i}[C_i + (N_i - 1).V_i] \\ N_{i+1} = N_i + 1 \end{cases}$$

Such a method converges quickly if we allow it to explore different behaviors. So in complement to the behavior driven by the previous equations, the agent is allowed to emit a random behavior with probability ϵ.

4.5 Q-Learning

Q-Learning is one of the most famous reinforcement learning method. Based on Sutton and Barto's Time Derivative (TD) model [28], this method has been proposed by Watkins [33]. Q-Learning is an algorithm for solving rapidly and easily stochastic optimal control problem. $Q_{s,a}$ represents the expected future payoff for action a in state s. Q-Learning works by modifing $Q_{s,a}$ for each pair of state-action using the following equation:

$$Q_{s,a} = Q_{s,a} + \alpha[r + \gamma \max_b Q_{s',a} - Q_{s,a}]$$

In a markovian environment and if the problem is stationary, it was shown that this algorithm converges with probability one to the optimal value. In practice, Q-Learning does not explore the environment sufficiently. Thus, it has been suggested to produce variability by adding noise. Contrary to the four other reinforcement learning methods presented above, Q-Learning is not "context-free": if a behavior is reinforced only in a precise situation, the animat will be able to emit this behavior in that situation only.

5 Results of simulation

Agents are tested in two situations. For the first(individual situation), they can choose between three behaviors which have direct consequences on their own counter. In the second situation (interactive situation), agents are put by couples in the same condition than human subjects. The algorithm performs a number of behavior that is rather similar as for human subject

In the individual situation (see columns 2, 3, and 4 of table 2), the environment is very simple and all algorithms succeed in adapting their behavior to optimize their rewards. The difference between algorithms is the speed of

	individual			social		
algorithm	R	P	N	R	P	N
law of effect	100	0	0	53	16	31
Hilgard and Bower	100	0	0	35	28	37
Mean earning	89	7	4	54	42	4
Q-Learning	89	7	4	58	17	25
Staddon-Zhang	66	0	34	86	1	13

Table 2. Distribution of different behaviors among a hundred agents after one thousand steps of simulation. R represents the behavior that provides reinforcements to its party, P is the bahavior that provides punishment and N is doing nothing. Reinforcements can be given to itself (individual situation) or to the other agent (social situation).

convergence. The only algorithm which exhibits a different behavior is Staddon-Zhang's law: within about thirty percent cases, it has a behavior which yields no consequence at all, which may seem a major weakness. However, a careful examination of the behavior of this algorithm shows that it keeps exploring its environment while all other algorithms have settled into a rest point.

In the "social" situation (see columns 5, 6, and 7 of table 2), results are opposite. By far, results of Staddon-Zhang are the best: at the end, it cooperates nearly nine times out of ten. Only Staddon-Zhang, the law of effect, and Q-Learning exhibit results that differs significantly from a random behavior. The very good results of the Staddon-Zhang's algorithm clearly show that the performance of a method designed for stationary problems can be very different in a dynamical situation. As said before, amongst the 5 algorithms, the algorithm relying on Staddon-Zhang's law has the unique feature to keep exploring its environment. This feature might be held as a weakness in a stationary environment, but it is the source of its strength in a dynamical situation. Cooperative rate are plotted on figure 1(b) and 1(d) at the beginning and at the end of simulation from random behavior, the algorithm comes to mostly cooperate.

6 Discussion and perspectives

By using a minimal social situation, we show how cooperation, is possible if the behavior of subjects is selected by their environment. The interest of such situations is twofolds. First, there can be a precise recording of emitted behaviors so that the analysis is possible. Second, they can be simulated by simple adaptive agents.

The minimal social situation used has shown that the asynchronization of agents' behavior is an important issue since synchronization implies cooperation straightforwardly with a high probability. So, the emergence of the synchronization of the agents' behavior is an important part of the route towards cooperation. More generally, it should be clear that the synchronization of actions, which generally lies implicitly in the background of lots of works dealing with simulations, is not a mere detail.

By using a *minimal social situation*, we may suppose that behavioral selection can be important to account for the emergence of cooperation. The use of adaptive agents in the same situation supports such an analysis. So, adaptive agents are a precious tool to test hypothesis about social behaviors. However, we emphasize that simulations never proove anything firmly. Reciprocally, experiments based on human or animal behavior provide insights into the design of agents. In many cases, animal behavior, as the result of million years of evolution, is able to optimize very complex situations. By knowing the mathematical relation between physical characteristics of environment and behaviors, we might be able to build a new generation of architectures of artificial agents, able to learn in many situation. This kind of agents should be able, for instance, to learn to imitate and to learn from verbal instruction [14,15].

This procedure — controlled social situation and simulation by agents — will be used in order to study more complex behaviors such as work division or sequential decision making in social situations. If the situation is more complex, many reinforcement learning algorithms used in this paper may become unusable because, apart for Q-Learning, they are "context free". Hence their behavior cannot be truly controlled by the characteristics of their environment, unless they are adapted to integrate the context. Due to its performance in a dynamical environment as well as its soundness with regards to the experimental analysis of behavior, we will work towards making Staddon-Zhang's law context sensitive.

References

[1] Attonaty, J.M., Chatelin, M.H., Garcia, F., and Ndiaye S.M., Using extended machine learning and simulation technics to design crop management stategies. *In EFITA First European Conference for Information Technology in Agriculture*, (1997) Copenhagen

[2] Axelrod, R., *The evolution of cooperation*, Basic Book Inc. (1984)

[3] Bergen, D.E., Hahn, J.K., Bock, P., An adaptive approch for reactive actor design, *Proc. European Conference on Artificial Life*, (1997).

[4] Boren, J.J., An experimental social relation between two monkeys, *Journal of the experimental analysis of behavior*, **9** (1966) 691–700.

[5] David S. Touretzky, Lisa M. Saksida, Skinnerbots, *Proc. 4th Int'l Conf. on Simulation of Adaptive Behavior, From Animals to Animats 4*, Maes, Mataric, Meyer, Pollack, Wilson (eds), MIT Press, 1996

[6] Dawkins, R., *The Selfish Gene*, Oxford University Press, Oxford (1976)

[7] Delepoulle S., Preux Ph., and Darcheville J.C., Partage des tâches et apprentissage par renforcement, *Proc. Journées Francophones d'Apprentissage*, (1998), 201–204 (in french)

[8] Deneubourg, J.L., and Goss S., Collective paterns and decision-making, *Ethology Ecology and Evolution*, **1** (1989) 295–311.

[9] Donahoe, J.W., Burgos, J.E., and Palmer, D.C., A selectionist approach to reinforcement, *Journal of the experimental analysis of behavior*, **60** (1993) 17–40.

[10] Hake, D.F., Vukelich, R., Analysis of the control exerted by a complex cooperation procedure, *Journal of the experimental analysis of behavior*, **19** (1973) 3–16.

[11] Hamilton, W. D., The Genetical Evolution of Social Behaviour, *Journal of Theoritical Biology* **7** (1964) 1–52.

[12] Hemelrijk, C.K., Cooperation without genes, games or cognition, *Proc. European Conference on Artificial Life*, (1997).

[13] Hilgard, E.R. and Bower, G.H.: *Theories of learning* (fourth edition). Prentice-Hall, Enblewood Cliffs, NJ.

[14] Hutchinson, W.R. Teaching an agent to speak and listen with understanding: Why and how? Proceedings of the Intelligent Information Agents Workshop, CIKM, Baltimore: http://www.cs.umbc.edu/ cikm/iia/submitted/viewing/whutchi.html

[15] Hutchinson, W.R., *The 7G operant behavior toolkit: Software and documentation*, Boulder, CO: Behavior System.

[16] Ito, A, How do selfish agents learn to cooperate? *Proc. Artificial life 5*, Langton, Shimohara (eds), MIT Press, (1996) 185–192.

[17] Kaelbling, L.P., Littman, M.L., and Moore, A.W., Reinforcement learning: a survey, *Journal of Artificial Intelligence Research*, **4** (1996) 237–285

[18] Langton Ch., Artificial life, *Proc. Artificial Life*, Langton (ed), Addison-Wesley, (1987), 1–47

[19] McFarland D., Towards Robot Cooperation, *Proc of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats 3*, (1994) 440–444.

[20] Murciano, A., Millán, J.R., Learning signaling behaviors and specialization in cooperative agents. *Adaptive Behavior*, **5(1)** (1997) 5–28

[21] Sidowski, J.B., Reward and Punishment in a Minimal Social Situation, *Journal of Experimental Psychology*, **55** (1957) 318–326.

[22] Sidowski, J. B., Wyckoff, B., and Tabory, L., The influence of reinforcement and punishment in a minimal socila situation. *Journal of Abnormal Social Psychology*, **52** (1956) 115–119.

[23] Skinner, B.F., *The behavior of organisms*, (1938). Englewood Cliffs, NJ: Prentice Hall.

[24] Skinner, B.F., *Science and human behavior*, (1953) New York: Macmillan.

[25] Skinner B.F., Selection by consequence, *Science*, **213**, 501–514, 1981

[26] Staddon, J.E.R., *Adaptive Behavior and Learning*, (1981) Cambridge University Press.

[27] Staddon, J.E.R., and Zhang, On the Assignment-of-Credit Problem in Operant Learning, *Neural Network model of Conditioning and Action*, M.L. Caumais S. Grossberg (eds), (1991) Laurence Erlbaum : Hillsdale, N V.

[28] Sutton, R.S. and Barto, A.G., "Time-Derivative Models of Pavlovian Reinforcement", *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, M.Gabriel and J.Moore (eds.), (1990) 497–537. MIT Press: ftp://ftp.cs.umass.edu/pub/anw/pub/sutton/sutton-barto90.ps

[29] Sutton, R.S., *Reinforcement Learning*, MIT Press (1998).

[30] Theraulaz G., Pratte M., and Gervet, J., Behavioural profiles in *Polistes dominulus* (Christ) wasp societies: a quantitative study. *Behaviour* **113**, (1990) 223–250.

[31] Thorndike, E.L., Animal Intelligence: An experimental study of the associative process in animals, *Psychology Monographs*, **2** (1898)

[32] Thorndike, E.L., *Animal Intelligence: Experimental studies.* (1911) New York : MacMillan.

[33] Watkins, C.J.C.H. and Dayan, P., Q-Learning. Technical Note. *Machine Learning*, **8(3)**, (1992), 279–292

[34] Wilson, E.O., *The Insect Societies.* (1971) The Belknap Press, Harvard University Press.