Selection of behaviors by their consequences in the human baby, software agents, and robots

Philippe Preux
Laboratoire d'Informatique du Littoral (LIL),
UPRES
Université du Littoral Côte d'Opale,
B.P. 719, 62228 Calais Cedex, France,
preux@lil.univ-littoral.fr
Samuel Delepoulle
URECA & LIL
delep@lil.univ-littoral.fr
Jean-Claude Darcheville
Unité de Recherche sur l'Évolution des Comportements
et des Apprentissages (URECA),
UPRES-EA 1059,
Université de Lille 3,
B.P. 149,
59653 Villeneuve d'Ascq Cedex, France,
darcheville@univ-lille3.fr

Abstract

Within a collaboration between computer scientists and psychologists, we are studying the acquisition and development of behaviors by animals, including human beings. The central hypothesis is that the behavior follows Thondike's law of effect (Thorndike, 1911) which indicates that the probability of emission of a behavior that is followed by favorable consequences increases. This law has largely been studied experimentally since then to explain and predict animal behavior. It is also known as the selection of behaviors by their consequences. We have implemented this law using reinforcement algorithms and we have been able to reproduce experiments involving human beings via simulations of reinforcement agents. Using this kind of techniques, we have recently been able to simulate the acquisition of an arm reaching movement. The simulation shows a remarkable similarity with the behavior of the human baby. It is able to reach different positions in the space and maintain these different movements. The movements, initially uncoordinated, become smooth and reach directly the target. In a close future, we will implement these methods in hardware robots.

1. Introduction

During dozens of years, computer scientists have considered that the implementation of a certain intelligence requires the implementation of cognitive processes and that the brain is a computer, that is something alike a Turing machine (as in other centuries, one saw Man like a clock [10]). These ideas were accompanied by a feeling that an analytical and reductionnist approach taking its roots in Descartes [7] was appropriate. Though still widely spread, there is now an alternative to the cognitive paradigm, which receives a growing interest. This alternative is grounded on a behaviorist approach in the sense that, rather than reproducing inaccessible mechanisms lying in the brain that are thought to produce intelligence, this approach aims at mimicing behaviors, and behavior dynamics [11]. Rather than passing Turing's test, some researchers put forward the idea of simulating rather modest organisms, such as insects, which have behavior abilities which are far in advance from those we are currently able to implement in artefacts. The analytical top-down approach is replaced by a bottom-up approach. This latter consists in putting in interaction a set of agents which are rather simple; from these interactions emerge interesting global dynamics. Today, the main drawback on this bottom-up approach lies in the fact that we are absolutely unable to predict the global dynamics of the set of interacting agents given the behavior of each agent, even though this behavior is perfectly well formalized and deterministic. This problem lies, at least in part, in our poor knowledge of non-linear dynamics. Emergent phenomena are ubiquitous: in physics, chemistry, biology, psychology, ethology, economy, politics, ... Despite these difficulties, researchers have performed the experiment: they have put a set of simple agents in interaction, just to see what happens (see [3] for example). Today, genetic algorithms, neural networks, ant colony algorithms, and, more generally, multi-agents systems solve real world problems, and exhibit interesting global dynamics, while we do not clearly understand how they give birth to them.

In this context, it is possible to give an account of the complex dynamics of a system if the behavior of its components is known with sufficient precision so that it can be implemented with algorithms. Then, the simulation of the system of agents should produce the expected global dynamics; otherwise, it is a sign that the supposed individual dynamics of agents is erroneous (or, more obviously, that it is not implemented correctly). In such a virtual laboratory, we can put to the test the consequences at the global level of the individual behavior of a set of interacting agents. The global dynamics resulting from the simulation can be compared to that of the original system in the real world. Then, we can participate to the study of living systems and provide a useful help to design reliable models of individual behaviors which do not break down when the agents of the

system are put to interact together. Indeed, in biology, ethology, or psychology, lots of models of an individual behavior simply do not fit any longer once they are put into interaction with such an other model with which it is supposed to interact in nature. This fact leads to a decomposition in different levels, such as behavior, individual, group of individuals, ... However, the reductionist approach can bring artificial problems. From the co-activity of agents can naturally emerge global dynamics of interest by self-organization. So, such a design of agents encompassing the ability to interact with other agents is interesting from two points of view. First, the model of behavior of an agent is richer; second, letting agents interact, self-organization plays its role and leads to the global dynamics. At this point, it is noteworthy that the global behavior of the system is than naturally much more adaptive than in a system designed at only one level.

Our work takes place in this context. Hence, we assume that the behavior of a living organism, or some parts of a living organism, is selected by its consequences, according to Thorndike's law of effect [15, 16]. This law indicates that the probability of emission of a behavior increases when its emission has been followed by some positive consequences. This law has been observed and measured in a full range of living organisms, ranging from flies to human beings. The principle of selection by consequences is also invoked to model the formation of groups of neurons in the cortex [9], as well as the way the immune system works. Though simple to express, we think with others that the interaction of agents which behavior is selected according to this law can give rise to very complex behavioral dynamics. This law provides important adaptive capacities to organisms. Then, we study the dynamics resulting from the interaction of agents which behavior is selected according to the law of effect. It is important to note that this system of agents is decentralized and unsupervised: it is decentralized because there is a set of agents acting concurrently; it is unsupervised in the sense that nothing tells the agents what to do and which goal they have to achieve. In our work, the objective is two-fold: showing that the behavioral dynamics of a living organism can be explained by this law; building artefacts that exhibit complex behaviors. We have put to the test this approach by studying the emergence of cooperation between adaptive agents [6, 4, 5]. Based on this work, in this paper, we show that a set of agents which behaviors are selected according to their consequences can explain the development of the reaching arm movement. The dynamics that is obtained shows striking similarities with that of the human new born. Then, we discuss the perspectives that are opened by this work to build software and hardware adaptive agents.

2. An arm as a set of cooperating components

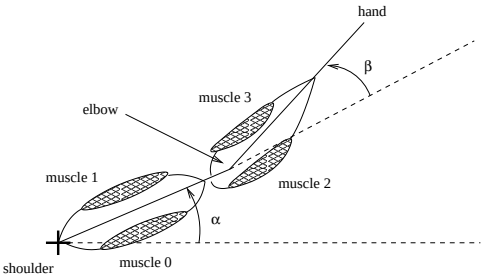

Figure 1: Our model of an arm: see text for explanations.

We consider an arm as a set of interacting agents and we wish it to develop an ability for a reaching movement, that is, putting its "hand" at a given location in the space. To this end, we have to precise the agents, their behavioral repertoire as well as the consequences of their behaviors. In the 2 dimensional model, an arm is made of two segments, each of which is controlled by two muscles (see fig. 1). One extremity of the arm is fixed: it is its shoulder; the other extremity can move under the action of the muscles: it is its hand. There are two articulations: the shoulder, and the elbow at the junction of the two segments. Two muscles control the rotation at the shoulder, the other two muscles control the rotation at the elbow. Each couple of muscles is made of one agonistic muscle and one antiagonistic muscle. We call this architecture MAABAC which stands for "Multi Agents Animat for Behavioral Arm Control".

Clearly, our point is not to model accurately the architecture of a real arm. A human arm contains hundreds of interacting muscles. We use a simplification which is compatible with works in neurosciences which show that muscles are acting within synergetical groups [12]. These simplifications are acceptable as far as we wish to demonstrate that a non centralized architecture may emit coordinated movements, rather similar to those observed in the reality. We are more interested in a phenomenological similarity than in a high accuracy in the architecture from which the behaviors originate.

Each muscle is in one out of N possible states of contraction. The position of a segment of the arm is related to the contraction of its two muscles. The angles α and β (see fig. 1) are obtained as follows: $\alpha = \frac{\pi}{2}\frac{c_0 - c_1}{N} + \frac{\pi}{2}$, and $\beta = \frac{\pi}{2}\frac{c_2 - c_3}{N} + \frac{\pi}{2} - \alpha$, where c_0, c_1, c_2, and c_3 are the contraction of muscles m_0, m_1, m_2 and m_3.

The task MAABAC has to fulfill is to put its hand in a certain zone of the space (this zone is called the reinforcement zone). Initially, MAABAC is not aware of this fact, and the operator can move the zone at will. Table 1 sketches the algorithm performed by MAABAC.

The direct consequence of a muscle contraction c is its energetic cost obtained by $\kappa(\frac{c}{N})^2$, with $\kappa \in]0, 1]$. Due to the opposite effect of their action, the equilibrium points for a couple of muscles are when both are in a state of contraction $c = \frac{N}{2}$. The second, and last, direct conse-

initialize $c_{i,0\leq i\leq 3}$ with a random number in $\mathcal{U}[0, N[$
for each time step loop
  for each muscle i
    select a behavior among : $\begin{cases} \text{contraction} : c_i \leftarrow c_i + 1 \\ \text{relaxation} : c_i \leftarrow c_i - 1 \\ \text{hold still} : c_i \leftarrow c_i \end{cases}$
  compute α and β
  compute the consequences of the behavior for each muscle $\kappa_{i,0\leq i\leq 3}$
  update the visual stimuli (d, a)
end loop

Table 1: Algorithm performed by MAABAC. See text for explanations.

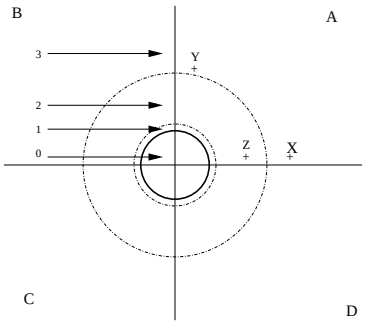

Figure 2: The visual field of MAABAC is split into 4 quadrants (A, B, C, and D) and 4 circular zones (0, 1, 2, and 3). Positions X and Z are perceived differently; positions X and Y are perceived as identical though more remote than X to Z. So, this visual system is rather crude and does not help very much the arm to reach the reinforcement zone.

quence is a visual stimulation. At this point, it is important to precise that what we have called a muscle up to this point actually models the system made of an organic muscle and its motor-neurons. Otherwise, it would not make sense to assume that the agent receives some sort of visual stimulus. So, a very poor visual system is modeled as follows: it provides 2 integer values (d, a), each in the range 0..3. d gives an indication of the distance between the hand and the center of the reinforcement zone while a gives an angular indication. Actually, the visual system can be switched off so as to make MAABAC blind. A very large majority of points of the space are merely perceived as far, in a certain quadrant (see fig. 2). This visual system has been designed to be very crude. There are two reasons for that: first, in the newborn, the visual system is also very crude. Second, from a computer scientist point of view, if the visual system was very acute, then, it would be very simple to solve the problem we have assigned the arm by simply moving the hand towards the target using a mere gradient algorithm. As far as the vision does not provide enough information to guide the hand towards the target, such a trick is not possible. This is precisely what we wanted to avoid: putting the solution of the problem into the agent.

There may also be an indirect consequence of the activity of the muscles: when MAABAC puts its hand in the reinforcement zone, it receives a reward $r = 1$; when the hand lies outside the reinforcement zone, the reward is $r = 0$. Then, each agent receives the following consequences after each time step: $r - \kappa(\frac{c}{N})^2$.

In the algorithm, the most delicate part is the selection of the behavior of each muscle. There, we want MAABAC to follow the law of effect. The problem is that the implementation of the law of effect is not obvious; there are several possibilities, some of which we discussed in [6] are based on reinforcement algorithm [14], while others are based on neural networks [8, 13]. To a certain extent, the reinforcement algorithm named Q-learning, introduced by Watkins [17], is a good choice, as argued by Barto [1]. Therefore, we have used Q-learning to implement the law of effect in the muscles of MAABAC.

According to the current state and its reinforcing value, this algorithm iteratively determines the next state. It also modifies the table of Q-values $Q_{s,b}$ which is used by the algorithm to predict the forthcoming reinforcement which it is likely to receive in a given state, when emitting a certain behavior. Let the emission of the behavior b in state s leads to the reinforcement r and the transition to state s', the modification of $Q_{s,b}$ is performed according to the following equation:

$$Q_{s,b} \leftarrow Q_{s,b} + \lambda(r + \gamma \max_a Q_{s',a} - Q_{s,b})$$

where λ and γ have a value in the $[0, 1[$ range. λ is the learning rate, while γ represents the relative importance of predicting the future reinforcements with regards to current reinforcement.

In MAABAC, the Q values are initialized to 0. The current state is defined by the intensity of the current contraction of the muscle as well as by the current perceptual stimuli. The behavior that is emitted is chosen to be the one with the highest predictive Q value with probability 0.9, at random otherwise; we have also used

Natural system	Implementation in MAABAC
Learning (law of effect)	Q-Learning
Group of sensori-motor neurons	Artificial agent
Coordination	Co-dependence of reinforcing agents
Structure of retina	Poor vision Variable density of receivers (higher in the center)
Contraction cost	Cost is a quadratic function of the contraction

Table 2: Comparison between a natural living system and the implementation of MAABAC.

a selection of behavior which is proportionate to their Q values using a Boltzmann selection. These two techniques basically aim at exploring the space of behaviors; this exploration is crucial for a good behavior of the algorithm.

Table 2 parallels the natural structures with their implementation in MAABAC.

3. Simulation and results

A simulator of MAABAC has been written in Java. The reinforcement zone is set at a certain position. The operator has the ability to set MAABAC arm in a certain initial position, and move the reinforcement zone at will. Then, the simulation is activated. The simulation of MAABAC shows different stages of development of the movement (see fig. 3): initially, the hand wanders erratically; after some times, it passes into the reinforcement zone, by chance, but it is unable to remain in it; after the position of the arm has been reset to its initial position by the operator, and after many attempts, the arm develops a smoother and smoother direct movement to reach the zone. Then, it is time to change the reinforcement zone. Of course, the arm will first seek the reinforcement zone in its former position. The thing is that as far as the arm will not receive any reward, it will wander again. Then, it will find the new location. However, it takes less time to reach the new zone than the first one. This fact shows what psychologists call a capacity of generalization. We can go on and on. Each time, the arm develops a smooth and direct movement. The more it has been trained on different positions of the reinforcement, the quicker the arm finds new positions. Finally, we can also reset the reinforcement zone to a former position. Then, the arm finds this former position very quickly. So, the arm has learnt many positions of the reinforcement zone and has developed the ability to reach it by smooth movements. We have also performed two types of experiments with MAABAC. One is known as "extinction" in the ethology literature; once a smooth movement has been learnt, extinction consists in suppressing all rewards, even if the zone is reached. After a while, the hand has no longer any reason to put its hand in the zone and wander erratically. Then, if we give the reward again (for the same position of the reinforcement zone), once the arm has passed along the zone, it shows a smooth and direct movement towards the zone again. So, this means that the movement has not been forgotten, but simply that, as long as it was no longer rewarded, it had disappeared temporarily.

The second kind of experiments we have done is that of shaping; in the behavior analysis literature, shaping is of utmost importance. Shaping is a technique that is used to teach a behavior to an animal that would not be emitted otherwise, though the animal is physically able to do it. Shaping goes as follows: instead of reinforcing the behavior that is sought and which occurrence has a very low probability, we first reward a certain behavior that is naturally emitted. Then, the rewarded behavior slowly drifts towards the unlikely behavior. At the end of this process, the originally unlikely behavior becomes "normal" for the animal (in other words, shaping can be called training: for example, a bear that rides a bicycle has been shaped to). Using shaping, we have been able to teach MAABAC to reach positions that it would not have reached otherwise, because they simply correspond to very unnatural positions of the arm.

4. Discussion

The performance of MAABAC shows that a non centralized and unsupervised system which dynamics is based on the law of effect is able to perform complex behaviors which are very comparable to those observed in animal and human babies. The learning curve of MAABAC is shown to demonstrate features that are typical of living organisms, such as human babies. It naturally supports shaping; it shows usual extinction of behaviors. Though not reported here, it is also able to track a mobile target. This being said, MAABAC is a simplification, almost a caricature, compared to the diversity of constraints handled by a physical system such as the arm of a baby. Thus, we will integrate problems related to gravity, add more segments, and their muscles, to the arm, as well as the third dimension. These add-ons require to tackle different sorts of questions related to physics, bio-mechanics, or neuro-physiology. This also implies to study the impact of the use of the control algorithms.

In our work, the objectives are two-fold. Concerning the study of dynamics of behavior of living beings,

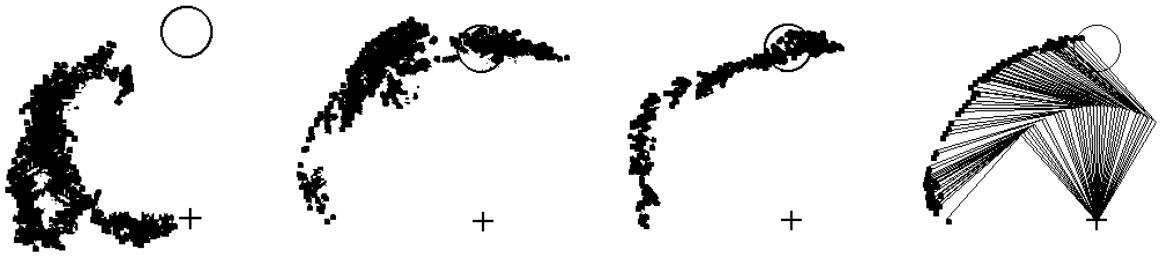

Figure 3: The development of the reaching movement: from left to right, the arm first has an erratic movement; after a while, it occasionally comes through the reinforcement zone; the movement becomes smoother and smoother; finally, the arm has a direct and smooth movement. The cross indicates the shoulder of the arm. The circle is the reinforcement zone. In the 3 leftmost plots, the position of the hand is indicated. In the rightmost plot, the whole arm is sketched.

this work argue for a selection-based evolution of some of animal behaviors. We want to stress an important point regarding our work on the arm reaching problem, a problem that have been tackled in many other works. In the classical simulation of the reaching movement [2], one uses the fully coordinated movement of the arm to write it down into equations of motion. To the opposite, we put the emphasis on the development phase, and we rely on a set of cooperating systems placed under a reinforcement contingency to learn the required movement.

Our other concern is that of building adaptive artefacts that are able to develop new behaviors in response to the contingencies laid by their environment. For us, this work is a step towards this direction. We are currently collaborating with a team of roboticists to obtain a hardware arm which would demonstrate the same abilities.

Acknowledgment

During this research, S. Delepoulle has been supported by contract n^o 97 53 0283 from "Conseil Régional Nord-Pas de Calais", France.

References

[1] A.G. Barto. *Reinforcement Learning and Adaptive Critic Methods*, pages 469–491. Van Nostrand Reinhold, 1992.

[2] N.E. Berthier. Learning to reach: A mathematical model. *Developmental Psychology*, 32:811–823, 1996.

[3] R.A. Brooks. *Cambrian Intelligence.* MIT Press, 1999.

[4] S. Delepoulle. *Coopération entre agents adaptatifs ; étude de la sélection des comportements sociaux, expérimentations et simulations.* PhD thesis, Université de Lille 3, URECA, Villeneuve d'Ascq, October 2000. Thèse de doctorat de Psychologie.

[5] S. Delepoulle, Ph. Preux, and J-C. Darcheville. Dynamics of temporal organization of behaviors dynamique in interaction situation, 2000. (submitted).

[6] S. Delepoulle, Ph. Preux, and J-C. Darcheville. Evolution of cooperation within a behavior-based perspective: confronting nature and animats. In *Artificial Evolution'99*, volume 1829 of *Lecture Notes in Computer Science*, pages 204–216. Springer-Verlag, 2000.

[7] R. Descartes. *Discours de la méthode.* J'ai lu, Paris, first edition in 1637.

[8] J.W. Donahoe, J.E. Burgos, and D.C. Palmer. A selectionist approach to reinforcement. *Journal of the experimental Analysis of Behavior*, 60:17–40, 1993.

[9] G.M. Edelman. *Neural Darwinism.* Basic Books, 1987.

[10] J. Offray De La Mettrie. *L'homme-machine.* new edition by Mille et une nuits, Paris, 2000, first edition in 1747.

[11] R. Pfeifer and C. Scheier. *Understanding Intelligence.* MIT Press, 1999.

[12] M. R. Rosenzweig and A. L. Leiman. *Physiological Psychology Second Edition.* Random House, Inc., 1989.

[13] K.R. Stephens and W. Hutchison. Behavorial personnal digital assistants: The seventh generation of computing. *The analysis of verbal behavior*, 10:149–156, 1992.

[14] R.S. Sutton and A.G. Barto. *Reinforcement Learning.* MIT Press, 1998.

[15] E.L. Thorndike. Animal intelligence: An experimental study of the associative process in animals. *Psychology Monographs*, 2, 1898.

[16] E.L. Thorndike. *Animal Intelligence: Experimental Studies.* Mac Millan, 1911.

[17] C.J.C.H. Watkins. *Learning from delayed rewards.* PhD thesis, King's college, Cambridge, UK, 1989.