

# Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation

**Gianluca Baldassarre**

Department of Computer Science,  
University of Essex,  
CO4 3SQ Colchester, United Kingdom  
gbalda@essex.ac.uk

**Domenico Parisi**

Institute of Psychology,  
National Research Council,  
Viale Marx 15, 00137 Rome, Italy  
parisi@ip.rm.cnr.it

## Abstract

Traditionally classical and instrumental conditioning have been studied in laboratory conditions in great detail but without trying to explain their role in organisms' adaptation. For example, little has been done to clarify how classical and instrumental conditioning mechanisms work together in an integrated fashion to enhance organisms' survival and reproductive chances. In this paper we argue that the adaptive role of classical and instrumental conditioning should be analysed in order to understand their ultimate meaning. One way to do so is to do simulations that on one side explicitly consider their adaptive function and on the other match at least some of the empirical data collected in the laboratory. We describe some simulations in which an organism learns to search for food in an attempt to clarify the role that some aspects of classical and instrumental conditioning may play in the development of this adaptive behaviour. We show how these two mechanisms work in an integrated fashion and how the model is validated by some psychological and neurophysiological data coming from the laboratory.

## 1. Introduction

The essence of classical conditioning can be illustrated using Pavlov's (1927) classic experiment. A dog is presented with a tone (conditioned stimulus) and subsequently with food (unconditioned stimulus). At the beginning of the experiment when food is presented the dog starts to salivate (conditioned response). After several trials the dog learns that food follows the tone, so it starts to salivate as it hears the sound.

Instrumental (or operant) conditioning can be described through an experiment made by Thorndike (1911). A cat is put in a closed cage from which it can see some food outside. In the cage there is a lever that opens the cage if pulled

down. The cat exhibits a sequence of actions chosen from its behavioural repertoire. Eventually it will pull down the lever and reach the food (instrumental response), obtaining the reward of food (reinforcement). If the experiment is repeated several times, the probability (per time unit) of producing the correct response tends to increase (law of effect), while the time taken to open the cage (latency) tends to diminish.

Since the experiments cited were done, classical conditioning and instrumental conditioning have been massively studied in laboratory conditions by psychologists. These studies have produced a great amount of important empirical data and the discovery of several related sub-phenomena analysed in great detail (e.g. extinction, relearning, conditioned inhibition, blocking, second order conditioning; cf. Lieberman, 1993). Now we know better *how* classical and instrumental conditioning work. Despite this, these studies have often failed to clarify *why* all these mechanisms are there: assuming an evolutionary point of view, what is their function in terms of organisms' adaptation to the environment? How do they increase the organisms' survival and reproduction chances? Also, the laboratory studies have usually treated the two mechanisms as independent, probably as a consequence of the aforementioned failure to consider their adaptive function. In this paper we argue that in order to have a complete picture of classical and instrumental learning we need to determine their role in organisms' adaptation. Furthermore, if we do so we will realise that classical and instrumental conditioning work together in a much more integrated fashion than the laboratory studies have shown. These arguments and the simulations presented in the paper can be viewed as an attempt to make explicit and develop some ideas proposed by Shultz et al. (1997), and Barto et al. (1990).

We think that the simulative research community can do much in building this wider picture. Among others Parisi et al. (1990) have shown that the simulation approach is particularly appropriate for studying how organisms adapt to the environment. In fact, in a computer simulation it is possible to simultaneously represent and study the reciprocal

interactions among an organism's "mind", its body, and the environment (ecological simulations). The capacity to learn plays a central role in more advanced species' adaptation because in the presence of environmental change it allows an organism to change its behaviour so to be more effective and efficient in avoiding danger, finding food, having sexual success, and taking care of offspring. It is becoming clear that classical and instrumental conditioning are at the very core of animals' learning (Rolls, 1999). Ecological simulations allow us to draw a general picture of how classical and instrumental conditioning underlie the evolutionary advantages of learning.

Empirical data from laboratory experiments play a crucial role in this picture. Some researchers have already addressed this data with specific computational models (for a review cf. Balkenius and Moren, 1998). These models are extremely simplified and directed at reproducing the specific details of the empirical data. They help us to understand the nature of the big variety of mechanisms that make up classical and instrumental learning but are limited in their scope. What we hope for is that empirical data coming from laboratory experiments were used to suggest and validate computational models of *complete* organisms with body, sensors, effectors, and "mind", that adapt to the environment on the basis of classical and instrumental conditioning.

In this paper we present a simulation within the framework we have illustrated. We simulate an embodied organism that learns to reach efficiently for some elements of food present in the environment. The organism's controller is largely based on the actor-critic model proposed by Barto et al. (1990) implemented with neural networks (cf. Lin, 1992). The simulation shows how classical and instrumental conditioning mechanisms allow the organisms to learn to collect food in a progressively decreasing number of steps, i.e. with reduced consumption of energy. As we shall see, the "critic" network of the model can be interpreted as an instantiation of (some of) the principles of classical conditioning, and the "actor" network as an instantiation of (some of) the principles of instrumental conditioning.

To validate the model we use two kinds of data, psychological and neurophysiological. The psychological data are the basic ones from the experiments of Pavlov and Thorndike mentioned at the beginning. The neurophysiological data are those of Shultz et al. (1997) synthetically illustrated in Figure 1. The authors describe the results of an experiment where single-neuron activity is recorded during the execution of classical conditioning experiments. The experiments were carried out with monkeys and involved the study of dopamine neurons in the midbrain. These neurons show a phasic activation (the proportion of neurons that fire is greater than in baseline activation) when monkeys are presented with various appetitive stimuli (small apple morsel, fruit juice, etc.). If the presentation of reward is repeatedly preceded by auditory or visual cues, the dopaminergic neurons change the time of the phasic activation from the time of reward to the time of cue onset. Surprisingly, if the cue is not followed by a reward anymore, the dopaminergic neurons show an average activation markedly below the

baseline activation at the time that reward should have occurred.

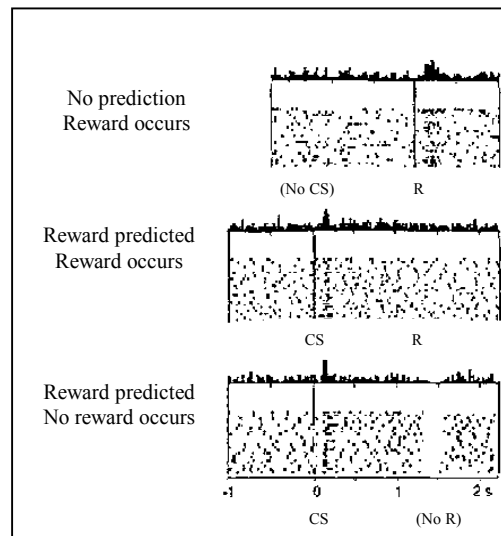


Figure 1: Each panel shows the time histogram of impulses from the same neuron. Each line of dots shows one trial. Horizontal distances of dots correspond to real-time intervals. CS: conditioned, reward-predicting stimulus. R: primary reward. (Reprinted with permission from Shultz et al. (1997). Copyright 1997, American Association for the Advancement of Science).

Section 2 of the article describes the environment and the organism's neural-network controller used in the simulations. Section 3 interprets the model in terms of classical and instrumental conditioning mechanisms. Section 4 describes the experiments and proposes an interpretation of the results. Section 5 draws the conclusions.

## 2. The environment, the organism, and the neural controller

The environment is a two-dimensional 1x1 unit square arena. In this arena there are 10 randomly distributed circular food elements, each with a 0.01 radius. The organism is a circle with a 0.02 radius (Figure 2). The simulation takes place in discrete time steps (input/output cycles of the neural network). If in one cycle the organism steps on a food element, it eats it and the food element disappears. When a food element is eaten a new element is introduced in a random location of the arena. The organism has an ingestion sensor that activates with 1 if in the current cycle a food element is ingested, with 0 otherwise. Also the organism has a one-dimensional "retina" of six non-overlapping sensors. These sensors receive information from a 180° frontal visual field. Each sensor has a scope of 30° and a depth limited to 0.1. A sensor activates with 1 if a food element, or part of it, is within its field, with 0 otherwise.

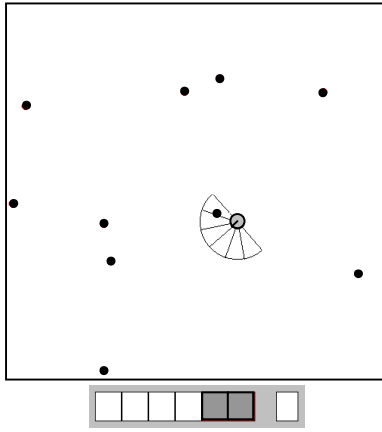


Figure 2: Top: The environment and the organism. The environment contains 10 food elements (small black circles) and one organism (big grey circle). The little line on the organism indicates the organism's current heading. The field covered by each of the six sensors is also shown. Bottom: Current activation of the 6 visual sensors and 1 ingestion sensor: only the 2 last visual sensors on the right (in grey) are active.

The organism has two legs (wheels). By controlling the length of the left and right step, the organism can go straight (same length for both left and right step) or turn (different lengths). The length of the left and right step can be either 0 or 0.02.

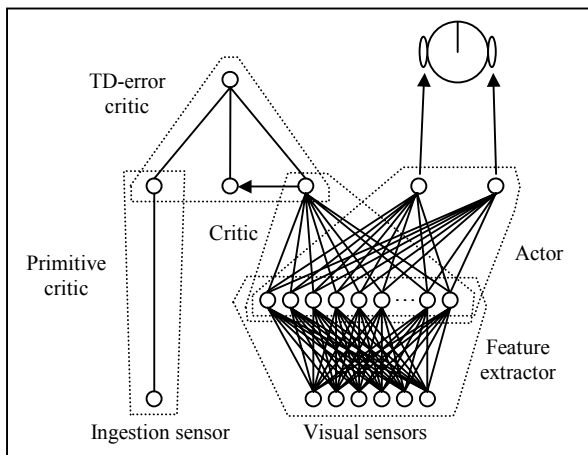


Figure 3: The components of the neural architecture controlling the organism's behaviour.

The architecture of the controller of the organism is represented in Figure 3. This figure divides the whole architecture of the neural controller in five parts: the feature extractor, the primitive critic, the TD-error critic (these three systems have fixed weights), the critic, and the actor (these two sub-systems have weights that change with learning). In Figure 4 we have represented the main aspects of the neural network.

*Feature extractor.* The feature extractor is based on the Kanerva coding (Kanerva, 1988; Sutton and Whitehead, 1993). Each input unit is connected with all the 30 output units (feature units). Both the  $n$  input units,  $x_i$ , and the  $m$

output units,  $y_j$ , assume a value in the set  $\{0, 1\}$ . Each weight,  $w_{ij}$ , is randomly chosen with equal probability in the set  $\{-1, +1\}$ , and is kept fixed during the simulations. The activation rule for a feature unit is:

$$y_j = \begin{cases} 1 & \text{if } \sum_{i=1}^n [w_{ji} x_i] \geq \vartheta_j \\ 0 & \text{else} \end{cases} \quad \vartheta_j = S_j^{\min} + \beta n$$

where  $\vartheta_j$  is a threshold, dependent on the feature unit,  $S_j^{\min}$  is minus the number of negative weights of unit  $y_j$ , and  $\beta$  is a constant in the interval  $(0, 1)$ , set to 0.6 in the simulations. According to this rule, each feature unit maximally responds to an input pattern called "prototype". The prototype for the unit  $j$  is the input pattern with  $x_i = 1$  for those  $i$  where  $w_{ij} = +1$  and with  $x_i = 0$  for those  $i$  where  $w_{ij} = -1$ . Because  $\beta = 0.6$ , the threshold  $\vartheta_j$  has a value such that the unit will be active ( $y_j = 1$ ) if more than 60% of the input bits match the prototype. The idea behind this rule of activation is that each feature unit activates with 1 only if the Hamming distance between the input pattern and the prototype is smaller than  $(1-\vartheta_j)$ .

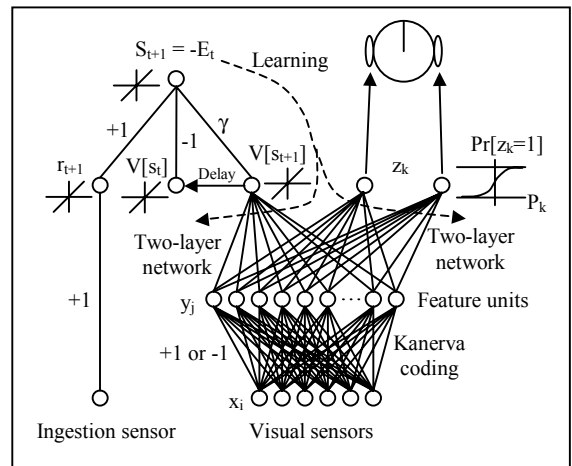


Figure 4: The main aspects of the neural network controlling the organism's behaviour.

The feature extractor based on the Kanerva coding has been introduced for two reasons. The first is to decrease the interference phenomena that occur with on-line learning (Sutton and Whitehead, 1993). In fact a feature unit will have an activation of 0 each time that the input pattern is too dissimilar from its prototype. In these cases the weights of the critic and the actor that correspond to the feature unit will not be changed by the Widrow-Hoff rule used in the training (see later). As a result they will specialise in giving an answer to the input patterns that are similar to its prototype, and will not be changed by other patterns. Furthermore, if the critic and the actor were directly connected with the sensors, they would not have enough degrees of freedom to give the proper answers. For example, in correspondence to the activation of two contiguous sensors the critic should necessarily (and probably wrongly) give an evaluation equal to the sum of the two evaluations given to the separate acti-

vation of the two sensors, since only one weight would be connected to one sensor.

*Actor.* The actor is a two-layer feed-forward network that takes the activation of the feature units as input and has 2 stochastic output units. The action potential  $p_k$  of the output unit  $k$  is computed as the weighted sum of the input signals ( $w_{kj}$  are the network's weights):

$$p_k = \sum_{j=1}^m [w_{kj} y_j]$$

Each stochastic output unit has an activation  $z_k$  in the set  $\{0, 1\}$ . The probability that it has an activation of 1, denoted with  $\Pr[z_k = 1]$ , is computed with the sigmoidal function applied to the activation potential:

$$\Pr[z_k = 1] = \frac{1}{1 + \exp[-p_k]}$$

The activation of the unit is determined on the basis of this probability as follows:

$$\begin{cases} z_k = 1 & \text{if } u \leq \Pr[z_k = 1] \\ z_k = 0 & \text{if } \Pr[z_k = 1] < u \end{cases}$$

where  $u$  is a random number drawn from a uniform distribution in the interval  $[0, 1]$ . The first output unit is used to determine the length of the left step. An activation of 0 corresponds to a step of length 0, an activation of 1 corresponds to a step of 0.02. The second output unit is used to determine the length of the right step.

*Primitive critic.* The primitive critic is a simple network that maps a signal coming from the external world, in our case the activation of the ingestion sensor, into an internal reward signal. In the simulations the weight of the connection between the ingestion sensor and the internal reward unit is fixed to +1, so the organism perceives a reward of +1 when an element of food is ingested, a reward of 0 otherwise. This network has to be thought of as innate.

*Critic.* The critic is a two-layer feed-forward network with one linear output unit (its activation is equal to the activation potential) which takes the activation of the feature units as input. During the simulation the critic has to learn to evaluate the current state of the world  $s_t$  by giving as output a signal  $V^\pi[s_t]$ . This signal is an estimation of the expected discounted sum of all future rewards that will be obtained starting from the current state  $s_t$  and using the action-selection policy  $\pi$  of the actor. In formal terms the (true) value  $V^\pi[s_t]$  that the critic has to learn to estimate, is defined as follows:

$$V^\pi[s_t] = E^\pi[\gamma^0 r_{t+1} + \gamma^1 r_{t+2} + \gamma^2 r_{t+3} + \dots]$$

where  $E^\pi[\cdot]$  denotes the function that returns the average value (recall that the action-selection policy of the actor is stochastic) given the current policy  $\pi$  of the actor,  $r_{t+1}$ ,  $r_{t+2}$ , etc., are the rewards at time  $t+1$ ,  $t+2$ , etc., and  $\gamma$  is a discount coefficient chosen in the interval  $[0, 1]$ . The choice between 0 and 1 of the coefficient  $\gamma$  has the consequence that the more the reward is far in the future, the less weight it receives. For example if  $\gamma = 0.95$ , as in the simulations, we have:  $\gamma^0 = 1$ ,  $\gamma^1 = 0.95$ ,  $\gamma^2 = 0.9$ ,  $\gamma^3 = 0.86$ ,  $\gamma^4 = 0.81$ , etc.

*TD-error critic.* This network implements in neural terms the computation of the TD-error (Temporal Difference error) of the TD learning method of Sutton (Sutton and Barto, 1998). Let us consider how the TD-error is computed in mathematical terms. On the basis of the same principles we have used to define  $V^\pi[s_t]$ , the (true) value  $V^\pi[s_{t+1}]$  associated with the state  $s_{t+1}$  is:

$$V^\pi[s_{t+1}] = E^\pi[\gamma^0 r_{t+2} + \gamma^1 r_{t+3} + \gamma^2 r_{t+4} + \dots]$$

On the basis of this formula,  $V^\pi[s_t]$  can be expressed in terms of  $V^\pi[s_{t+1}]$ :

$$\begin{aligned} V^\pi[s_t] &= E^\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots] \\ V^\pi[s_t] &= E^\pi[r_{t+1} + \gamma (r_{t+2} + \gamma^1 r_{t+3} + \dots)] \\ V^\pi[s_t] &= E^\pi[r_{t+1}] + \gamma E^\pi[r_{t+2} + \gamma^1 r_{t+3} + \dots] \\ V^\pi[s_t] &= E^\pi[r_{t+1}] + \gamma V^\pi[s_{t+1}] \end{aligned}$$

This equation is represented graphically in Figure 5.

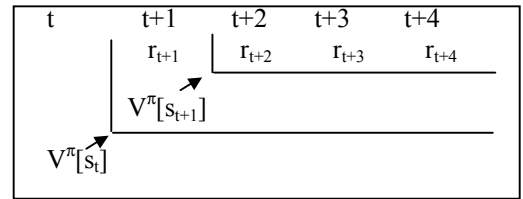


Figure 5: Graph that shows how the evaluation made at time  $t$  is equal to the evaluation made at time  $t+1$  plus the reward obtained at time  $t+1$  (instead of the average values the effective rewards  $r$  are considered).

If we pass from the true values  $V^\pi[s_t]$  and  $V^\pi[s_{t+1}]$  to the estimations given by the critic and affected by errors,  $V^\pi[s_t]$  and  $V^\pi[s_{t+1}]$ , the last equation does not hold and we have:

$$V^\pi[s_t] \neq r_{t+1} + \gamma V^\pi[s_{t+1}]$$

where we have substituted the average value  $E^\pi[r_{t+1}]$  with the specific value  $r_{t+1}$  observed at time  $t+1$ . The two sides of the equation can be considered as two different estimations of the same value  $V^\pi[s_t]$  associated with the state  $s_t$  and made at time  $t$  (left side) and  $t+1$  (right side) respectively. The estimation on the right side is built as a sum of the estimation  $V^\pi[s_{t+1}]$  of  $s_{t+1}$  made by the critic at time  $t+1$  (and weighted with  $\gamma$ ) plus the reward effectively observed at that time. The right-side estimation, even if not correct because of the uncertainty of the component  $V^\pi[s_{t+1}]$ , tends to be more correct than the estimation on the left side. In fact its component  $r_{t+1}$  is directly observed, so it is accurate. For this reason the difference  $E_t$  between the two estimations:

$$E_t = V^\pi[s_t] - (r_{t+1} + \gamma V^\pi[s_{t+1}])$$

can be considered as a proxy of the error of estimation  $V^\pi[s_t]$  made by the critic at time  $t$ . We also define:

$$S_{t+1} = -E_t = (r_{t+1} + \gamma V^\pi[s_{t+1}]) - V^\pi[s_t]$$

as the "surprise" of the TD-critic at time  $t+1$ . Notice that in the literature on TD-methods what we have just defined as surprise  $S_{t+1}$  is usually called "error". An explicit definition of surprise is useful to interpret the results of our simulations. The TD-critic network computes the difference be-

tween  $V^\pi[s_t]$  and  $(r_{t+1} + \gamma V^\pi[s_{t+1}])$ . The comparison of these two values can result in either a confirmation of the expectation (the old evaluation  $V^\pi[s_t]$ , in which case the difference is 0) or a positive/negative surprise. The reason to represent in neural terms the computation of the TD-error is that the activation of the surprise unit ( $S_{t+1}$ ) has a parallel in the activation of the dopaminergic neurons studied by Shultz et al. (1997). Also these authors have proposed a neural implementation of the TD-error computation.

*Learning of critic.* The critic is trained with a Widrow-Hoff algorithm (Widrow and Hoff, 1960) that uses as error the error signal coming from the TD-critic. In particular the weights  $w_j$  of the output unit are changed so that the estimation  $V^\pi[s_t]$  tends to be increasingly close to the more accurate target value  $(r_{t+1} + \gamma V^\pi[s_{t+1}])$ :

$$w_j = -\eta \left( V^\pi[s_t] - (r_{t+1} + \gamma V^\pi[s_{t+1}]) \right) y_j = -\eta E_t y_j = \eta S_{t+1} y_j$$

where  $\eta$  is a learning rate set to 0.05 in the simulations. Given an actor with certain weights that expresses a particular (stochastic) policy  $\pi$ , the critic will learn to produce an increasingly precise estimation  $V^\pi[s_t]$  of the true  $V^\pi[s_t]$ .

*Learning of Actor.* The actor is trained with a modified version of the algorithm proposed by Ackley and Littman (1991). If the surprise is positive, the weights of the actor are changed so that if the same input is met in the future, the action  $(z_1, z_2)$  has a higher probability of being selected. This is accomplished by changing the weights so that each of the probabilities  $\Pr[z_k=1]$  moves toward the target value  $z_k$  correspondent to the action currently selected:

$$w_{jk} = -(\zeta S_{t+1}) (\Pr[z_k = 1] - z_k) (\Pr[z_k = 1] \Pr[z_k = 0]) y_j$$

where all the values, with the exception of the surprise  $S_{t+1}$ , refer to time  $t$ . Notice that this is again a Widrow-Hoff algorithm where  $(\zeta S_{t+1})$  is the learning rate made proportional to the surprise ( $\zeta$  is 0.05 in the simulations),  $(\Pr[z_k = 1] - z_k)$  is the error, that moves the probability toward its desired target  $z_k$ ,  $(\Pr[z_k=1]\Pr[z_k=0])$  is the value of the derivative of the sigmoidal function calculated on the point  $p_k$ , and  $y_j$  is the activation of the feature unit. If the surprise of the critic is negative, the weights are not updated. In this case the desired action is unknown, and changing the weights so that the probabilities approach the complement to 1 of the output pattern (as in Ackley and Littman, 1991) did not have much effect.

To understand the logic underlying the learning of the actor, suppose that the critic is very efficient and instantaneously learns to produce a correct evaluation of the current state  $s_t$ . In this case the critic's estimation  $V^\pi[s_t]$  corresponds to the discounted sum of the future rewards that will be obtained on the average starting from  $s_t$  and acting according to the actor's policy  $\pi$ . One of the reasons that  $V^\pi[s_t]$  is an average is because the action that the actor selects in correspondence to a given state is stochastic. Another reason is because if the world is stochastic the consequences of actions are stochastic. Now suppose that the actor selects an action that produces a positive surprise in the critic. This means that on the average this action is better than the ac-

tions that the actor tends to produce in correspondence to the same input. So this action deserves to be strengthened in terms of probabilities of being produced when the same input pattern is met in the future. In the simulations the learning of the critic and of the actor take place contextually (policy iteration). The critic learns to predict with increasing accuracy what the consequences (in terms of future rewards) of the actions selected by the actor are. At the same time the actor learns to strengthen all the actions that (positively) surprise the critic when selected. This should lead the actor to an increasingly better policy.

### 3. Interpretation of the model in terms of classical and instrumental conditioning

In the model the neural network implementing the "critic" is an instantiation of the classical conditioning mechanisms. This network learns to assign an evaluation to each successive perceived state of the world according to how promising it is in terms of delivery of reward and punishment in the future. The surprise calculated in correspondence to unexpected events that follow the evaluation, is considered to be the neural event that underlies the conditioned response in classical conditioning experiments. In the simulations the surprise signal is at the very core of the learning process leading the critic's network to produce the correct evaluation and the actor's network to express an adaptive behaviour. The conditioned response, like the salivation of Pavlov's dog, is not explicitly present in the model because in general it serves specific adaptive functions (it helps the digestion in the case of the salivation) but is not essential for the role that classical conditioning plays in learning. In the classic experiments the conditioned response was important because it was a measurable physiological indicator of the internal event "surprise", otherwise not observable. In experiments like the ones of Shultz et al. (1997) the surprise is directly measured at the neural level. In the model the unconditioned stimulus is the food ingested. In general the unconditioned stimulus is the ultimate event/object that increases the organism's fitness. In the model the conditioned stimulus is each stimulus that precedes the consumption of food. Notice that in ecological conditions we usually have a sequence of relevant stimuli as opposed to a single special one.

The neural network implementing the "actor" is an instantiation of the instrumental conditioning principles, and the locus where the adaptive behaviour is developed and stored. Before learning this network expresses the same probability of producing each of the actions of the organism's repertoire (with random weights the output of the sigmoidal units is close to 0.5). During learning the network's weights change so to yield a higher and higher probability for the actions that have led to high and unexpected rewards. In the model the "instrumental response" is the behaviour expressed by the organism, i.e. the sequence of actions that lead to food. The "discriminating stimulus(i)" is the sequence of stimuli preceding the consumption of food. The

"reward" is the activation (+1) of the primitive critic's output unit due to the ingestion of food.

Notice that the same (visual) stimuli perceived by the organism function both as the unconditioned stimuli of classical conditioning and as the discriminating stimuli of instrumental conditioning. Similarly the activation of the ingestion sensor caused by the food represents both the unconditioned stimulus of classical conditioning and the direct cause of the reward perception of instrumental conditioning.

## 4. Experiments and results

We have organised the simulations in three groups, each with a different experimental condition.

### 4.1 Simulation 1: No learning

In the first simulation the experimental condition was that the actor and the critic had initial random weights and did not learn. The measured dependent variables were the evaluation  $V$  of the critic, the surprise  $S$  and the error  $E$ . The value of these variables was measured for a sequence of 10 cycles divided in 5 sub-sequences:

1. 2 cycles before the cycle in which an element of food enters the visual field.
2. 1 cycle in which an element of food enters the visual field.
3. 4 cycles before the organism steps on the food element. If the organism took more than 4 cycles from seeing the food to reaching it, the data relative to the central cycles were not considered. For example, in the case of 5 cycles the 3<sup>rd</sup> cycle was eliminated. In the cases in which the organism took just 3 cycles, the 2<sup>nd</sup> cycle was considered two times.
4. 1 cycle in which the ingestion of food takes place.
5. 2 cycles following the ingestion of food.

Each simulation was repeated with 9 different random seeds (hence 9 different organisms). For each random seed, two ingestions of food were considered. Ingestions of food involving the presence of more than one element of food in the visual field in one of the 10 cycles were not considered. The values of the dependent variables for each of the 10 steps were averaged across the 18 ingestions of food. The results of the simulations are shown in Figure 6. This picture plots the evaluation of the critic and the error and the surprise of the TD-error critic, measured in correspondence to the five stages of the total sequence of 10 cycles.

As it could be seen on the computer screen, the organism's behaviour was erratic and only occasionally led to food ingestion (for a quantification of this behaviour, see Figure 7, condition "without learning"). Figure 6 shows that given that the critic had not undergone learning, the evaluation  $V^m[s_t]$  associated to each state was always around 0, with the food inside or outside the visual field. For this reason when the food was eaten at cycle 8, the surprise bounced to about 1 since the reward  $r = 1$  was completely unex-

pected. For the same reason the evaluation at cycle 7 showed an underestimation error of about -1.

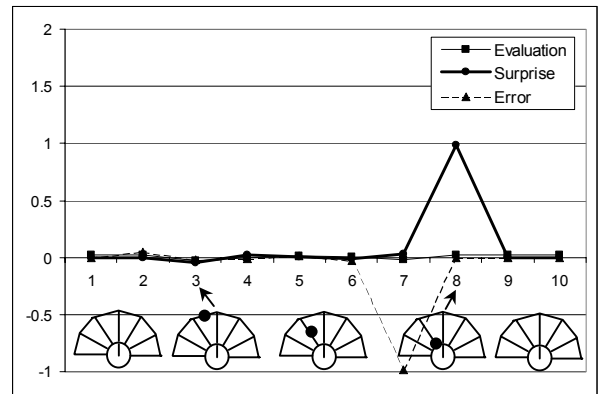


Figure 6: Evaluation of the critic, and surprise and error of the TD-error critic, measured for 10 cycles before, during, and after the ingestion of food. Experimental condition: random initial weights, no learning.

The shape of the plot of the surprise is qualitatively similar to the activation of the dopaminergic neurons recorded by Shultz et al. (1997) and shown in figure 1 (top graph). The differences are due to the fact that in our model the surprise is expressed by one neuron with a continuous activation, while in the brain it is expressed by the average activation of (a population of) spiking neurons.

### 4.2 Simulation 2: Learning

In the second group of simulations the experimental condition was that the actor and the critic had initial random weights but learning took place for 100,000 cycles. Then learning was ceased and some measures were taken. The effect of learning on the adaptation of behaviour could be observed on the screen. If no food was seen, the organism went quite straight at maximum speed, exploring different zones of the arena. When the food entered the visual field, the organism turned toward it and reached for it. Figure 7 gives a quantitative measure of the learning that occurred. In this figure we report the measure of the ability to search for food along the 100,000 cycles, averaged for 9 simulations (using the same random seeds that were used in the first group of simulations). For comparison the figure also reports the performance of the same 9 organisms without learning taking place (previous simulation). The ability to search for food was measured as the mobile average of the number of cycles taken by the organism to reach a food element. The average was based on the last 100 food elements reached. At the beginning of the simulation, when less than 100 elements of food had been reached, a simple average was measured. The strong irregularity of the performance of the untrained organisms and the "lucky" start of the trained organisms depend on the great variability of the chance to encounter food by moving randomly. The variability is stronger at the beginning given the small size of the window of the moving average. For example, the cycles needed by

the 9 untrained organisms to reach the first food element were 47, 16, 48, 231, 431, 104, 87, 529, 22.

From the graph it appears that learning had the effect of reducing the number of cycles needed by the organism to reach a food element from about 175 cycles (average performance without learning) to about 24 cycles. In a real organism this would mean a great reduction of the energy spent to search for food, and a big increase in the survival and reproduction chances. The graph is also consistent with the latency graphs of instrumental conditioning studied in laboratory (cf. Thorndike, 1911).

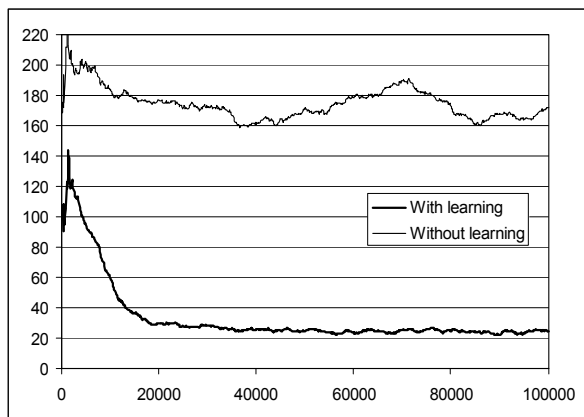


Figure 7: Average performance of 9 organisms, with and without learning, during 100,000 cycles.

After 100,000 cycles learning was stopped and the experimental dependent variables were measured according to the same protocol employed in simulation 1. The plot of them is shown in Figure 8.

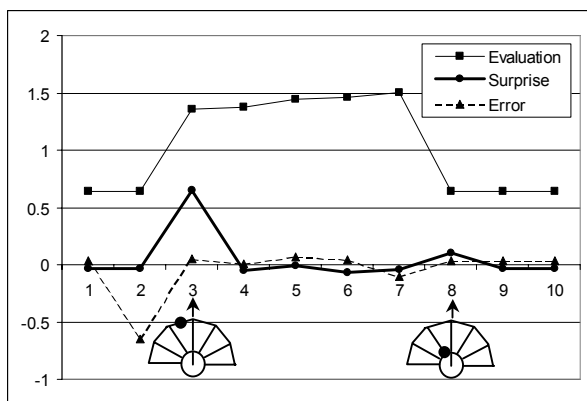


Figure 8: Evaluation of the critic, and surprise and error of the TD-error critic, measured after 100,000 cycles of learning.

This figure shows several interesting facts. The evaluation of the critic is above 0 (about 0.6) even if the food is out of sight. This happens because the critic has learned that moving in the environment implies a certain probability of bumping into a food element. When the food enters the visual field at cycle 3, the critic's evaluation jumps to a much higher level. This is due to the fact that the state in which the food is perceived is more promising, in terms of reward,

than the states without food, especially if the actor has learned to move efficiently to the food. When the food is ingested, at cycle 8, it's no more in sight, so the evaluation drops again.

The most interesting fact is that after learning the positive surprise moves from the moment of ingestion of food to the moment when food is seen the first time. This reflects the inner nature of classical conditioning. When the food enters the visual field, the critic correctly starts to emit a high evaluation of the current state. The "goodness" of these states could not be predicted in the previous cycles and this causes an underestimation error of the critic (cycle 2) and positive surprise of the TD-Critic when food appears (cycle 3). Also notice that error and surprise are close to zero in all the following steps, showing that the critic is able to correctly evaluate all the corresponding states so the TD-critic "is not surprised".

Again, the plot of the surprise matches the activation of the dopaminergic neurons shown in Figure 1 (middle graph). These results can be thought of as matching the data of Pavlov's behavioural experiments if these experiments were carried out in ecological conditions. To see why, let us assume that the organism of the simulation had a salivation system that was triggered by both the direct perception of reward (food in the mouth) and by a high critic's evaluation (the critic's evaluation is high when the ingestion of food is imminent). This would be advantageous in terms of adaptation because the salivation would start as soon as the organism anticipated the coming of the food thereby facilitating its digestion. Now if we were to measure the salivation of the organism of the simulations, we would make the following observations. Before learning the salivation would start with the consumption of food. After learning the salivation would start as soon as the food is seen and its ingestion anticipated. This would match the data coming from the psychological experiments of Pavlov, but in ecological conditions (with the difference that the conditioned stimulus would be the direct sight of food and not the sound of the bell). Measuring the critic's evaluation and the TD-error critic's surprise (as we have actually done in the simulation) would give an idea of the neural processes underlying this pattern of the salivation process: after learning the surprise, causing the salivation, would move from the moment when food is ingested, to the moment when food is seen. As we have seen, this is also what Shultz et al. (1997) have observed in their experiments.

What is the effect of the simultaneous learning of both the critic and actor on the action selection policy adopted by the actor? Measuring the actor's probabilities of selecting different actions with different sensory stimuli can clarify the nature of this effect. Figures 9 and 10 plot these measures before and after learning (for simplicity the probability of staying still is not plotted, and only a few combinations of the activation of sensors are considered). In order to calculate the probability of selecting a given action (e.g. go left) the joint probability of selecting the appropriate left step and the appropriate right step was calculated. For example, if the probability of selecting a left step of 0 was 0.8 and the prob-

ability of selecting a right step of 0.02 was 0.5, the joint probability of going left was  $0.8 \times 0.5 = 0.4$ .

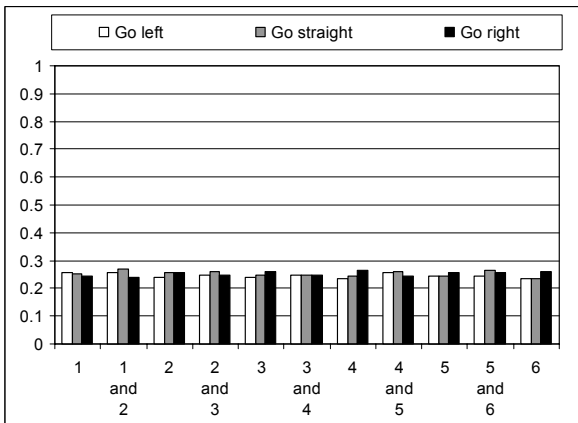


Figure 9: The probabilities of selecting different actions (ordinate) with different sensory stimuli (abscissa: the sensors with an activation of 1 are indicated) before learning.

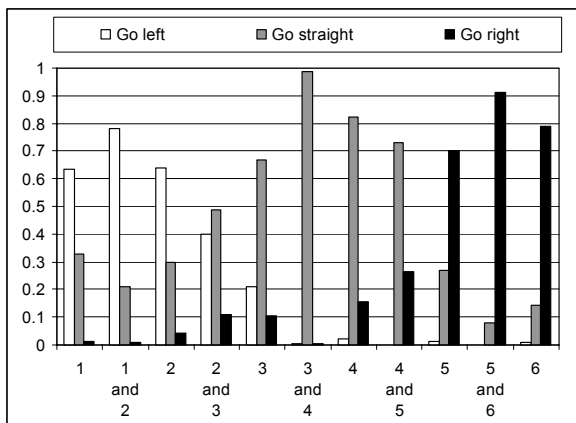


Figure 10: The probabilities of selecting different actions with different sensory stimuli after learning.

Figure 9 shows that before learning the actor has the same probability of selecting any one of the actions of its repertoire (about 0.25 for each action). Figure 10 shows that after learning the probability of selecting "go left", "go straight", and "go right", is very high when the food is on the left, in front, or on the right of the organism, respectively. These findings match the behavioural data resulting from experiments on instrumental learning, and in particular Thorndike's (1911) "law of effect".

### 4.3 Simulation 3: Food disappears

In the third group of simulations the experimental condition was the following one. The actor and the critic had initial random weights. Learning took place for 100,000 cycles. At this point learning was stopped and food disappeared as soon as the organism stepped on it and before it ate it. The effect was that the ingestion sensor and the visual sensors had an activation of 0. Under these conditions the experimental dependent variables were measured according to the

usual protocol. The results of the simulations are shown in Figure 11.

The evaluation of the critic is similar to the one of the preceding condition, and so are the error and the surprise of the TD-error critic until cycle 6. However, when the organism reaches the food at cycle 8, the reward is missing and this results in a negative surprise of about -1 of the TD-critic. Also the critic "did not expect" the drop of the evaluation in state 8 (due to lack of reward), so in state 7 there is an overestimation error of about +1.

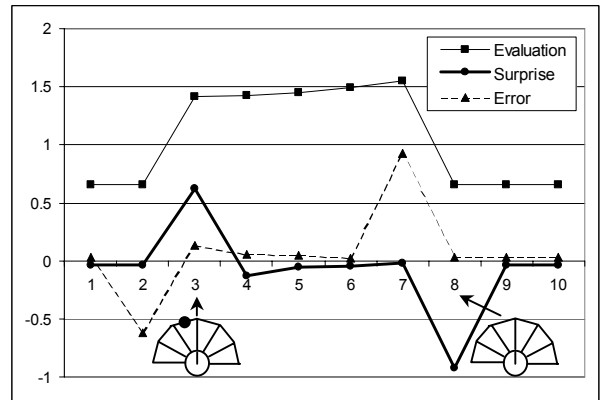


Figure 11: Evaluation of the critic, and surprise and error of the TD-error critic, measured after learning takes place for 100,000 cycles. Food disappears as soon as it is reached by the organism and before the organism eats it.

Notice that, again, the plot of the surprise matches the activation of the dopaminergic neurons shown in Figure 1 (bottom graph). The signal of the negative surprise has a role in the "extinction" phenomenon observed in the laboratory experiments on classical conditioning. "Extinction" consists in the disappearance of the conditioned response when the unconditioned stimulus fails to follow the conditioned stimulus for some trials. In ecological conditions this mechanism is adaptive because it allows the organisms to abandon behaviours that do not lead to a reward anymore. The model analysed here is capable of reproducing the extinction phenomenon, but this is not considered here.

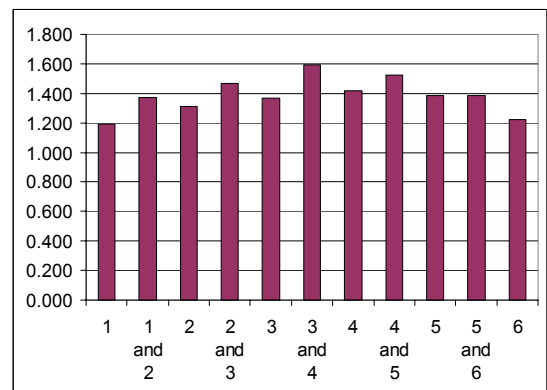


Figure 12: The evaluation of the critic (ordinate) in correspondence to the activation of different sensors and different couples of contiguous sensors (abscissa). Average for 9 organisms.

Figures 8 and 11 show another interesting fact. From step 3 to 7 the evaluation of the critic progressively increases. To explain this, let us refer to Figure 12. This figure shows that a lower evaluation is associated to the activation of the peripheral sensors versus the central sensors, and of one sensor versus two contiguous sensors. This happens because if food is perceived in a lateral position (peripheral sensors) or distant position (one sensor instead of two), on the average the organism takes more cycles to reach for it. More cycles imply that the reward associated with the food is strongly discounted (cf. the computation of  $V^m[s_i]$ ). When food appears it is likely to be in a lateral or distant position, so the evaluation of the early cycles of Figure 8 and 11 versus the later ones tends to be low.

## 5. Conclusion

The aim of the present paper was to stimulate a widening of the perspective with which classical and instrumental conditioning tend to be traditionally approached. The study of these phenomena with laboratory experiments and computational simulations has usually focused on understanding and reproducing the details of the two mechanisms, without a serious attempt to build a deep comprehension of their role in organisms' adaptation. For this reason it has also failed to enlighten the possible connections between the two mechanisms. We have argued that "ecological simulations" allow us to fill in this gap. Ecological simulations allow us to analyse with a single, integrated computational theory/model (the simulation) the possible interactions between the organisms and their environment. This makes it possible to study the role that classical and instrumental conditioning play in organisms' adaptation. Also we have argued that the data collected in laboratory should be used to validate these models.

To support our arguments we have presented some ecological simulations where an organism learns to reach for food with a progressively lower consumption of time (energy). At the basis of the learning mechanisms used in the simulations there is the concept of surprise, probably the core of classical conditioning, and the increase in the probability of actions that lead to situations that are highly promising in terms of reward, probably the core of instrumental conditioning. The model has been validated with some psychological and neurophysiological data. The simulations illustrate how classical and instrumental conditioning mechanisms work in close connection to augment the adaptation of the organism to its environment.

## Acknowledgements

We thank Prof. Jim Doran of the Department of Computer Science, University of Essex, for his valuable suggestions, and the same department for funding the first author's research.

## References

- Ackley D.H. and Littman M.L. (1991). Interaction between learning and evolution. In Langton C.G., Farmer J.D., Rasmussen S., and Taylor C. (Eds.). *Artificial Life II*. Reading - MA, Addison Wesley.
- Balkenius C. and Moren J. (1998). Computational models of classical conditioning: a comparative study. In Pfeifer R., Blumberg B., Meyer J.A., and Wilson S.W. (Eds.). *From Animals to Animats 5: Proceedings of the Fifth International Conference of Adaptive Behaviour*, pp. 348-353. Cambridge - MA, MIT Press.
- Barto A.G., Sutton R.S., and Watkins C.J.C.H. (1990). Learning in sequential decision making. In Gabriel M. and Moore J. (Eds.). *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. Cambridge - MA, MIT Press.
- Kanerva P. (1988). *Sparse Distributed Memory*. Cambridge - MA, MIT Press.
- Lieberman D.A. (1993). *Learning - Behaviour and Cognition*. Pacific Grove - CA, Brooks/Cole.
- Lin L.J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*. Vol. 8, pp. 293-321.
- Parisi D., Cecconi F., and Nolfi S. (1990). Econets: Neural networks that learn in an environment. *Network*. Vol. 1(2), pp. 149-168.
- Pavlov I.V. (1927). *Conditioned reflexes*. Oxford, Oxford University Press.
- Rescorla R.A. and Wagner A.R. (1972). A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and non reinforcement. In Black A.H. and Prokasy W.F. (Eds.). *Classical Conditioning II: Current Research and Theory*. New York, Appleton Century Crofts.
- Rolls E.T. (1999). *The Brain and Emotion*. Oxford, Oxford University Press.
- Shultz W., Dayan P., and Montague P.R. (1997). A neural substrate of prediction and reward. *Science*. Vol. 275, pp. 1593-1599.
- Sutton R.S. and Barto A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge - MA, MIT Press.
- Sutton R.S. and Whitehead S.D. (1993). Online learning with random representations. *Proceedings of the Tenth International Conference on Machine Learning*, pp. 314-321. Los Altos - CA, Morgan Kaufmann.
- Thorndike E.L. (1911). *Animal Intelligence*. New York, MacMillan.
- Widrow B. and Hoff M.E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*. Part IV, pp. 96-104.