

# Classical and Operant Conditioning as Roots of Interaction for Robots

Jean Marc Salotti and Florent Lepretre

Laboratoire EA487 Cognition et Facteurs Humains, Institut de Cognitique, Université de Bordeaux, 146 Rue Léo Saignat, 33076 Bordeaux Cedex, [salotti@idc.u-bordeaux2.fr](mailto:salotti@idc.u-bordeaux2.fr)  
69121 Heidelberg, Germany

**Abstract.** We believe that classical and operant conditionings play a fundamental role in the learning of adaptive behaviors in varying environments. We present a new model that integrates both conditioning mechanisms. It is based on a prediction system with traces of all events.

**Keywords:** Classical conditioning, operant conditioning.

## 1 Introduction

In the animal domain, classical conditioning, also called Pavlov conditioning, is a key mechanism in the learning of adaptive behaviors [6]. Operant conditioning is a complementary process, which enables animals training by means of associations between actions and rewards [10]. We believe that these processes are at the root of any complex interaction between animals and their environments, especially varying ones as it is expected when other cognitive agents are present. Interestingly, there has been a lot of work in the domain of reinforcement learning, but the results obtained in robot training are still very far from what can be obtained with animals [4] [12]. Models for classical and operant conditioning in the animal domain are therefore very interesting sources of inspiration to control the basic interaction mechanisms of robots. In this paper, we propose a new model for conditioning, which integrate classical and operant conditioning in the same system. In the next part, we recall some important properties of classical conditioning and related phenomena. Then, we present our model, which has been implemented and tested in a simulator.

## 2 Classical and operant conditioning

In the basic experience of classical conditioning, there is an unconditioned stimulus (US), a conditioned stimulus (CS), and a response R [6]. If the conditioning is successful, the specific response R should be observed whatever the presence or absence of the US. An important parameter is the time interval between the presentation of the CS and the presentation of the US (ISI=Inter Stimuli Interval). The

conditioning is strong and fast for very small ISI and it becomes more and more difficult as the value of the ISI increases. The timing of the US (reward or punishment) therefore plays an important role in the reinforcement process.

Other interesting behaviors have been observed when subtle variations are introduced in conditioning experiments. Latent inhibition occurs when the CS is presented alone several times before the standard conditioning protocol. The conditioning is still observed but the CS-US association should be repeated a greater number of times to obtain the same results. Latent inhibition is a key mechanism to avoid untimely associations. In experimental psychology, schizophrenia is typically considered as a consequence of a default in the latent inhibition process [1].

A "blocking" of a CS2-US association occurs when a CS1-US conditioning has already been performed and CS1 is always presented before CS2. Indeed, the logic behind that behavior is that CS1 is probably the true cause of the reward and CS2 is only a border effect. The CS2-US reinforcement is thus not justified in that case.

"Extinction" of the conditioning occurs if the CS is repeatedly presented without the US. However, if one tries to restart the conditioning trials after extinction, the reacquisition of the conditioning is faster than in the first place. Secondary conditioning occurs when a first CS (CS1) is used for classical conditioning and a second CS (CS2) is introduced before CS1. CS2 predicts CS1 and finally becomes a predictor for the US. The response is therefore observed when CS2 alone is presented.

In operant conditioning experiments, the animal should learn the consequence of its action. It is similar to classical conditioning, apart the fact that the conditioned stimulus is replaced by a conditioned action [10].

### 3 Our model

#### 3.1 Rescorla and Wagner model and reinforcement learning

Since the synthesis of experimental results presented by Pavlov, there has been a lot of work in that domain and different models have been presented. Most of them are based on the original model proposed by Rescorla and Wagner [2], [3], [5], [7], [9], [11]. Equation (1) gives the modification of the associative strength of a given stimulus X after a new trial. The increase is proportional to the salience of X (parameter  $\alpha$ ) and the efficiency of conditioning (parameter  $\beta$ ).  $\lambda$  is the maximum strength and  $V_{Total}$  is the sum of all associative strength of the present stimuli.

$$V_X^{n+1} = V_X^n + \alpha_X \beta (\lambda - V_{Total}^n) \quad (1)$$

The associative strength of a given stimulus can be interpreted as the degree of prediction of the US. Let us consider an example. In a basic conditioning experiment with a given stimulus X, the sum of all associative strength is equal to the associative strength of X and if the experience is repeated, its value converges towards  $\lambda$ . Suppose now that another conditioning with a stimulus Y is performed and confirmed before the conditioning with X. The sum of all associative strength becomes equal to

the associative strength of Y, which is close to  $\lambda$ . The second term is therefore close to zero and the associative strength of X does not increase much. This is how the blocking effect is explained by the model. Its ability to explain such a complex behavior is probably one of the key reasons of its popularity. However, it does not explain other important behaviors and the dynamics of the conditioning are not considered.

Sutton and Barto established the basic principles of reinforcement learning [12]. They also proposed a temporal difference model of classical conditioning, but it does not integrate operant conditioning [11]. Schmajuk, Lam and Gray proposed an interesting method that takes into account the novelty of the stimulus. It impacts on the attentional system and can explain latent inhibition [9]. However, despite this abundant literature there is still some debate on the exact processes explaining all aspects of classical conditioning [2]. For instance, in a recent paper, Rescorla discussed the mechanisms of spontaneous recovery, which is according to him still not well understood [8].

Fundamentally, the theory of reinforcement learning provides the key tools to implement classical and operant conditioning. However, we should pay attention to specific problems that are generally not fully addressed:

- Conditioning occurs in varying environments and all life long.
- The dynamics of the reward play an important role. The presence of a reward or a punishment is important but the efficiency of the reinforcement depends on its timing. As a consequence, time discretization has a significant impact and the state "reward after X seconds" should be distinguished from the state "reward after Y seconds".
- The representation of stimuli also plays an important role. For instance, suppose that a light is switched on and a reward is given after 10 minutes. The light is still on a few seconds before the reward, but the conditioning is impossible. Switching the light on is a possible stimulus that can be used in conditioning experiments, not the fact that the light is on. On the other hand, an increase or decrease of the frequency of a metronome can be used for conditioning. A stimulus is therefore not always specified by a precise event. The solution is probably to consider that conditioning is based on internal events corresponding to modifications in the representation of the world.
- Past experience plays an important role. In conditioning, everything is evolving and the system never comes back to a previous state.

### 3.2 Proposal

We propose an event-based model that is inspired from the works of Rescorla and Wagner. Each associative strength between two stimuli A and B is correlated with the degree of prediction that B should follow A. We therefore have to specify the mechanisms of a prediction system. A Bayesian network is typically appropriate. However, we would like to focus on the states that predict a reward or a punishment and we are not interested by the value of a specific descriptive variable but the exact time of its modification (the fact that the light is on does not matter, what matters is the event associated to the switching). In our Bayesian network, a state is thus defined by a representation of the world in terms of recent events. The value of a transition between state A and state B determines the probability that the event associated to

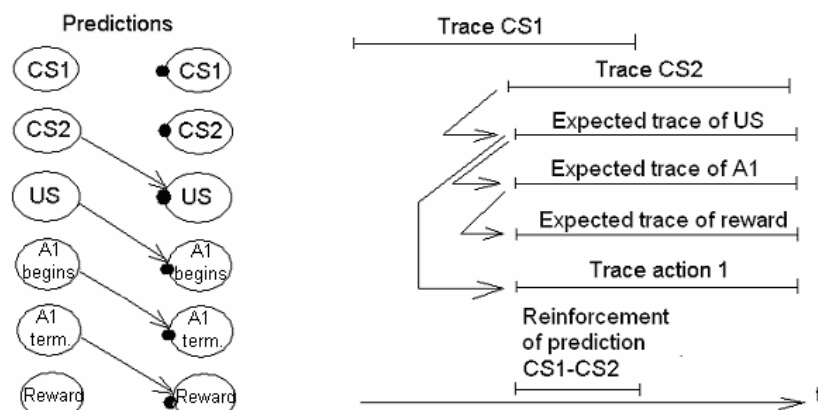
state B occurs after the event associated to state A. Moreover, that occurrence is constrained by a limited period of time. We propose using a trace of the event so that the stimulus is memorized a few seconds (we used 5 seconds in our experiments).

In order to integrate classical and operant conditioning, every action is specified by two states, the first one for the "action begins" event and the second for the "action terminates" event. Fixed transitions have been added in our network so that "action begins" always predicts "action terminates". In addition, if an action is a predictor of a given state S, we impose the transition to be defined between "action terminates" and S. And if a state S is a predictor of an action, we impose the transition to be defined between S and "action begins". Other transitions are prohibited.

Finally, it is also possible to consider that the states correspond to the neurons of a neural network and the values of the transitions correspond to synaptic weights.

### 3.2 Algorithm

We propose using the prediction system in a recursive way. See Figure 1. Given a state S, if it predicts another state (if the prediction value is greater than a threshold), the trace of an "expected state" is immediately activated and if it predicts other states, they also are activated. If in the sequence of states (real or expected) a reward is predicted, there is reinforcement. The reinforcement is simply an update of the prediction between two states. Our algorithm is presented Figure 2.



**Fig.1** : Example of application. The current predictions of the network are presented in the left part of the figure. CS1 does not predict anything, but CS2 predicts an US, which in turn predicts a specific action that leads to a reward. In the right part, the events of a trial are presented. CS2 follows CS1. As soon as CS2 is encountered, the US, a specific action and a reward are expected. The action is therefore immediately started and reinforcement between CS1 and CS2 occurs.

It is important to notice that in the same trial, the update might occur several times depending on time discretization and the overlap period of the traces. The closer the events and the greater number of times the value of the transition is increased. That

property is interesting since it is well known that animal conditioning is faster when the time interval between the CS and the US is shorter.

In our algorithm, reinforcement occurs even if the reward is only expected but finally not concretized. In that case, a punishment should be *a posteriori* performed. The problem is to determine the transition that has to be penalized. A "mistaken stimulus" has to be identified. In order to determine it, the sequence of all states that take part in a prediction of a reward is systematically memorized. As a first approach, we propose to select the state of the sequence with minimum transition value and to penalize it. Indeed, a stimulus present in the sequence for the first time is probably linked to the cause of the missing reward. The weakest value corresponding to the weakest causal link, it seems reasonable to penalize it. Such a strategy enables complex behaviors such as inhibitory conditioning.

The final term of the algorithm enables latent inhibition. If no reward was expected during the trace of a given state A, the salience of A is decreased. Such a mechanism is comparable to the attentional deficit proposed by Schmajuk *et al* [9].

```

For each active trace  $A_x$ 
  If reward present or expected
  For each active trace  $A_y$  ( $Y$  different from  $X$ )
    If (not Expected( $\dot{Y}$ )) or ( $W_{i \rightarrow j} > \text{threshold}$ )
      If  $\text{Start}(X) < \text{Start}(Y)$ 
         $W_{t+1}^{X \rightarrow Y} = W_t^{X \rightarrow Y} + \alpha_X \beta_Y (R_t - W_t^{X \rightarrow Y})$ 
For each finishing trace  $A_x$ 
  If  $A_x$  predicted a reward that did not occur
    Look for the set of possible mistaken stimuli
    For each mistaken stimulus  $M$  and prediction  $Y$ 
      If  $W_t^{M \rightarrow Y} > -1$        $W_{t+1}^{M \rightarrow Y} = W_t^{M \rightarrow Y} - \alpha'_M$ 
  If  $A_x$  occurred without expectation of any reward
     $\alpha_X = \alpha_X * \mu$ 

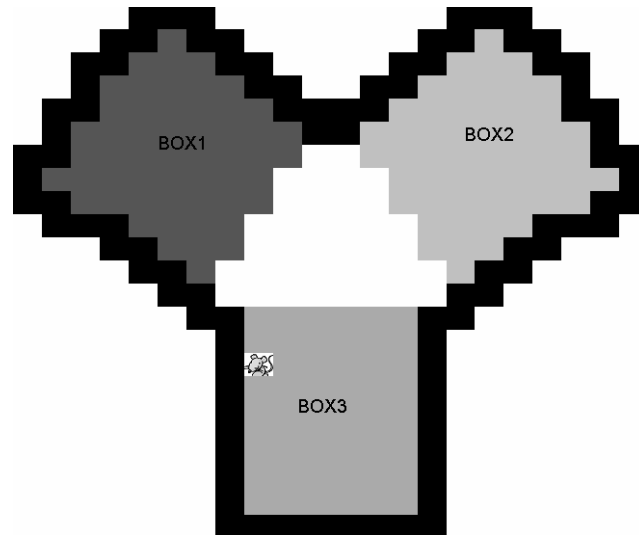
```

**Fig. 2:** Reinforcement algorithm.  $W$  is the degree of prediction,  $R$  the reward,  $\alpha$  is the salience of the stimulus,  $\beta$  is the efficiency of reinforcement,  $\alpha'$  is also linked to the salience of the stimulus but used for the penalty,  $\mu$  is the salience penalty (smaller than one).

## 4 Tests

### 4.1 Environment and possible events

Our model has been implemented in an autonomous agent. It determines the behavior of an artificial rat placed in a box with three compartments, see Figure 3. The environment is defined by a grid of 24x24 squares. Each square is a possible location for the animal except for those that are labeled "obstacle". A timer is used to control the behavior of the animal and enable real-time moves.



**Fig. 3:** Environment for testing behaviors. The only obstacles are at the frontier (painted black).

Every 0.3 seconds, stimuli are detected, events are processed and an action is decided. The animal can move one square at a time, horizontally, vertically or diagonally. Possible events associated to stimuli are:

- Hearing a bell.
- Seeing a light.
- Seeing and smelling food.
- Touching food.

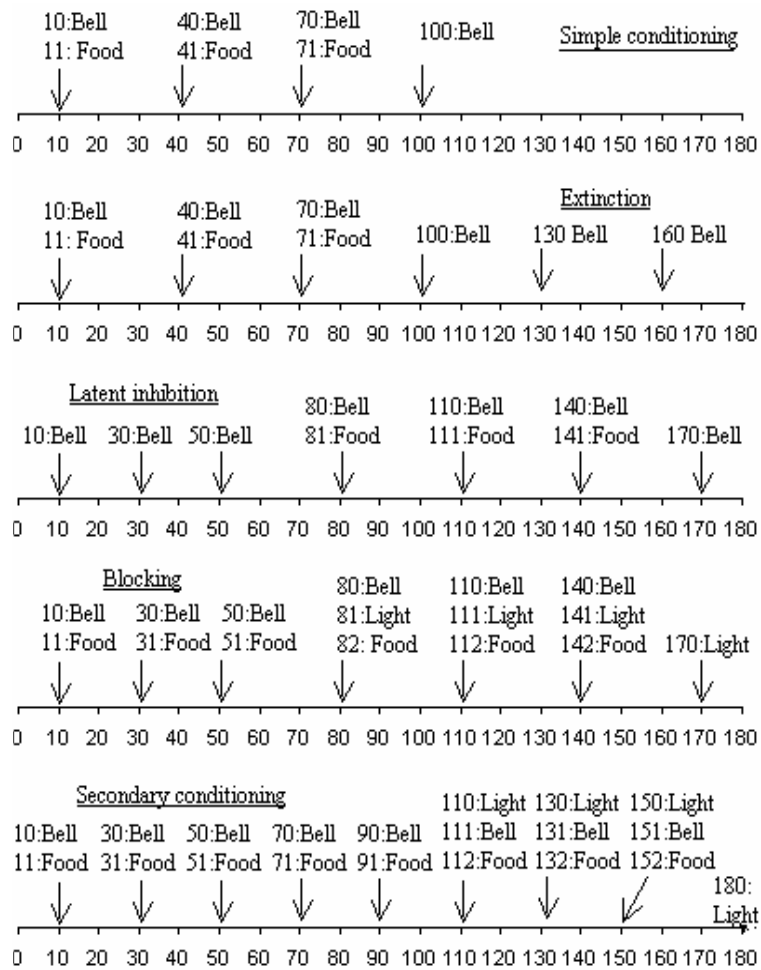
Other events are associated to actions:

- A1 starts: Start going towards food location.
- A1 terminates: Arrived at food position.
- A2 starts: Start going towards random position in Box 1.
- A2 terminates: Entering box 1.
- A3 starts: Start going towards random position in Box 2.
- A3 terminates: Entering box 2.
- A4 starts: Start going towards random position in Box 3.
- A4 terminates: Entering box 3.
- A5 starts: Start eating.
- A5 terminates: Stop eating.

In addition, the exact coordinates of the food are stored and updated in the "spatial representation" part of the system each time the food is set in a given square. Going to a specific location is made possible by a simple mechanism. All squares of the environment are duplicated in the spatial representation. Each one of the 8 neighboring squares is a candidate for the next move. The chosen square is simply the one that minimizes the distance to the target. Our environment is very simple but we wanted to focus on conditioning problems.

## 4.2 Scenarios

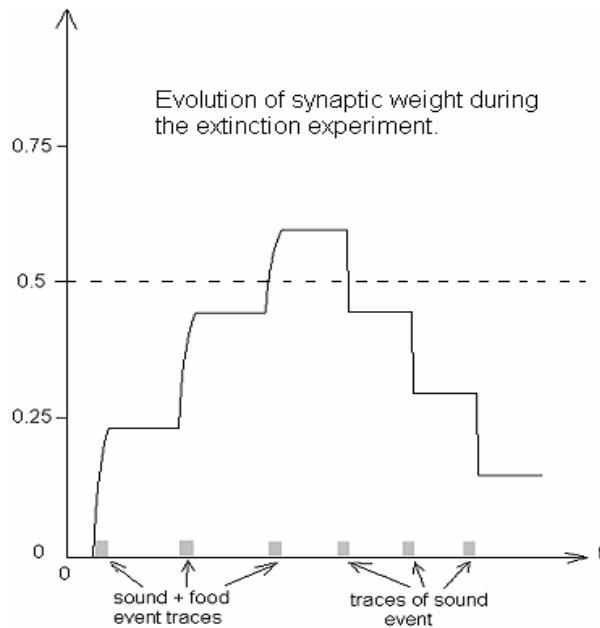
In order to test the model, specific scenario files have been elaborated, in which all environmental events are *a priori* defined. In order to avoid long tests, we imposed that simple conditioning between a CS and an US should occur after three trials with small ISI. The constants used in the reinforcement model have therefore been set to speed up the conditioning. Furthermore, each trial is separated from the previous one by at least twenty seconds. Five scenarios are presented Figure 4.



**Fig.4:** Different conditioning experiments have been conducted. From top to bottom: Simple conditioning, extinction, latent inhibition, blocking and secondary conditioning. Every environmental event is reported on the graph. All experiments last less than 180 seconds.

From top to bottom:

- In our experiment of simple conditioning, it is expected that three trials are sufficient. When hearing the bell after the three trials, if the conditioning is a success, the animal should start going to food location even if the food is absent.
- Extinction of conditioning should be observed if the bell is rung three times without presence of the food.
- Latent inhibition has a retarding effect on conditioning. In this case, three positive trials are considered not sufficient for conditioning. If the bell is rung, the animal should not start going to food location.
- If the bell is a predictor of the food, typically after three positive trials, no other stimulus coming after the bell can be associated with the incoming of the food. The bell is "blocking" the conditioning. In our experiment, it is expected that the light at the end of the sequence will not trigger the action of going to food location.
- A secondary conditioning effect can be observed if there is firstly a strong conditioning with the bell and the food (we used five trials) followed by a conditioning with a light switched on just before the ringing of the bell. At the end of the test, the expected behavior is the animal going to the food location as soon as he sees the light.



**Fig. 5:** Evolution of the associative strength between sound and food detection stimuli during the extinction experiment. Conditioning is effective if the weight exceeds 0.5. Parameters: traces last 5 seconds; timing 0.3 seconds;  $\alpha=0.001$ ;  $\beta=1$ ;  $R=1$ ;  $\alpha'=0.15$ ;  $\mu=0.9$

The correct expected behaviors have been observed in the five tests presented Figure 4. The evolution of the synaptic weight (or transition value) between the sound

detection state ("hearing a bell") and the food detection state ("seeing and smelling food") is presented Figure 5.

Operant conditioning has been tested separately. The user can add food in box 1 by pressing a button. If the button is systematically pressed when the animal ends its action "going to box 2" (or "going to box 3"), operant conditioning occurs. After three trials, as soon as the animal comes to the state "stop eating" (because everything has been eaten), it goes towards the box that predicts more food and comes back to eat it.

#### **4. Conclusion**

We got good results in the main experiments of conditioning. However, there are still some small problems in very specific cases occurring in more complex experiments that have not been presented here. For instance, when there is a sequence of several actions with an expected reward but finally no reward is obtained, the strategy used to penalize the weakest transition is not appropriate. The way the blocking phenomenon is obtained is also questionable. In our algorithm, the blocking is due to a comparison between the dates of the events before applying reinforcement. The blocking is therefore strict. In the Rescorla and Wagner model and in many other models, it is possible to significantly increase the associative strength between the intercalated CS and the US after a great number of trials. The blocking is therefore not total.

Another important issue is the efficiency of secondary and higher order conditioning. Our algorithm enables fast higher order conditioning thanks to the recursive projection system, which looks for the expected reward after a long sequence of expected events. In particular, it does not matter that the reward comes after several long actions combined with specific stimuli, what only matters is the fact that a reward is expected or not. Compared to other methods (see for instance the comparison made by Balkenius [2]), the efficiency of our method for higher order conditioning is interesting.

Finally, we also integrate in the same system classical conditioning and operant conditioning and this is another important contribution.

The model has also been implemented and tested with other stimuli and other actions in another simulator in order to prepare experiments with a real robot. We hope presenting more complete results in a near future.

#### **References**

1. Alvarez, R., De la Casa, L. and Sánchez, P., Latent Inhibition as a Model of Schizophrenia: from Learning to Psychopathology, *International Journal of Psychology and Psychological Therapy*, Vol. 3, N° 2, pp. 251-266, (2003).
2. Balkenius C. and Morén J., Computational models of classical conditioning: a comparative study, in Mayer, J.-A. , Roitblat, H. L., Wilson, S. W., and Blumberg, B. (Eds.), *From Animals to Animats 5*. Cambridge, MA: MIT Press, (1998).
3. Klopff, A., A neuronal model of classical conditioning, *Psychobiology*, vol. 16 (2), 85–125, (1988).

4. Lungarella, M., Metta, G., Pfeifer R. and Sandini G., Developmental robotics: a survey, *Connection Science*, vol. 15 (4), pp. 151-190, (2003).
5. Mignault, A. and Marley, A., A Real-Time Neuronal Model of Classical Conditioning, *Adaptive Behavior*, vol. 6 (1), pp. 3-61, (1997).
6. Pavlov, I.P., *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (translated by G. V. Anrep). London: Oxford University Press, (1927).
7. Rescorla R.A. and Wagner A.R., A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, In Black, A. H., & Prokasy, W. F. (Eds.), *Classical conditioning II: Current research and theory*, 64-99, New York: Appleton-Century-Crofts, (1972).
8. Rescorla, R.A., Spontaneous Recovery, *Learning Memory*, vol. 11, pp. 501-509, (2004).
9. Schmajuk, N.A., Lam, Y. and Gray, J.A., Latent inhibition :A neural network approach, *Journal of Experimental Psychology : Animal Behavior Processes*, 22 (3), pp. 321-349, (1996).
10. Skinner, B.F., *Science and Human Behavior*. New York: Macmillan, (1953).
11. Sutton R.S. and Barto A.G., A temporal-difference model of classical conditioning, *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, pp. 355-378, (1987).
12. Sutton R.S. and Barto A.G., *Reinforcement Learning: An Introduction*, MIT Press, (1998).