Prisoner's Dilemma Is Learned by Operant Conditioning Mechanisms

Instituto de Ingeniería Biomédica
Facultad de Ingeniería
Universidad de Buenos Aires
and
Instituto de Biología y Medicina Experimental—CONICET
silvano@fi.uba.ar

Abstract The prisoner's dilemma (PD) is the leading metaphor for the evolution of cooperative behavior in populations of selfish agents. Although cooperation in the iterated prisoner's dilemma (IPD) has been studied for over twenty years, most of this research has been focused on strategies that involve nonlearned behavior. Another approach is to suppose that players' selection of the preferred reply might be enforced in the same way as feeding animals track the best way to feed in changing nonstationary environments. Learning mechanisms such as operant conditioning enable animals to acquire relevant characteristics of their environment in order to get reinforcements and to avoid punishments. In this study, the role of operant conditioning in the learning of cooperation was evaluated in the PD. We found that operant mechanisms allow the learning of IPD play against other strategies. When random moves are allowed in the game, the operant learning model showed low sensitivity. On the basis of this evidence, it is suggested that operant learning might be involved in reciprocal altruism.

Keywords
Operant learning, neural networks, prisoner's dilemma, reciprocal altruism, game theory, reinforcement learning

1 Introduction

Natural selection can be conceived as a struggle for life in which only those living organisms that are best adapted to existing conditions and to changing environments are able to survive and, fundamentally, to reproduce. From this point of view, one of the most important apparent paradoxes of evolutionary theory is cooperation among individuals. In *The Origin of Species*, Darwin [15] noted that cooperation in social insects was a special difficulty which at first appeared to be insurmountable and fatal to the whole theory of evolution. However, he outlined a possible solution, stating that the difficulty disappears if we realize that selection may be applied to the family as well as to the individual.

Hamilton [24] in 1964 solved the paradox theoretically. There are groups of animals that sacrifice their own reproduction in collaborating with the reproduction of others that are related genetically. He found that these groups leave on average more copies of their own genes than noncollaborative ones. In this way, he concluded that there is selection at the family level. This process has been called kin selection; it explains, from the perspective of evolutionary theory, the existence of sterile castes in social insects such as bees, wasps, and ants. There are other theories to explain the evolution of cooperative behavior [17]: trait group selection [70], by-product mutualism [13], and reciprocal altruism [64, 65].

Table 1. PD payoff matrix: points earned by each player (A, B).

	Player A	
	Cooperate (C)	Defect (D)
Player B		
Cooperate (C)	(R=3, R=3)	(S=0, T=5)
Defect (D)	(T=5, S=0)	(P=1, P=1)

Reciprocal altruism involves the exchange of a benefit among two or more interactors and an associated cost of each altruistic act. If no individual fails in the reciprocity, in a long-term interaction each participant will experience a net benefit. This mechanism of cooperation will be selected if there is protection against lack of reciprocity in the altruistic act [5]. Using game theoretical analysis, Trivers [64] was the first to formalize these principal traits. He analyzed a nonzero-sum game, the well-known prisoner's dilemma (PD). The PD is the leading metaphor for the evolution of cooperative behavior in populations of selfish agents [38].

The PD has been studied in diverse areas such as evolutionary biology, sociology, philosophy, and economics. In the game, a player may either cooperate (C) or defect (D), and each player receives a payoff defined by Table 1. Generically we can establish the following equations so that the game maintains its characteristics:

$$T > R > P > S \tag{1}$$

$$2R > T + S \tag{2}$$

Two players have adopted a *Nash equilibrium* if each is playing a strategy that is the best reply to the other's strategy. Thus, if the PD is played only once, the only Nash equilibrium is defection, because it pays better regardless of what the opponent chooses. As a result, both players obtain 1 point instead of the 3 they would have obtained if they had chosen to cooperate.

In the *iterated prisoner's dilemma* (IPD), players face their opponents repeatedly. In order to maximize the payoff, the players can change their moves according to the opponent's strategy. If there is a fixed, known number of interactions between a pair of players, defecting is still the only strategy that is evolutionarily stable. But if the number of interactions is not fixed in advance, there is no single best strategy regardless of the behavior of the opponent [5].

In the evolutionary prisoner's dilemma (EPD), the scores that each player has obtained in an IPD are used to simulate the evolution of a population. Each new generation has a different arrangement of players proportional to the total score obtained against all adversaries. Thus, the total score for each strategy changes in each new generation, because the proportion of opponents changes.

Trivers was the first to relate reciprocal altruism to the IPD. Axelrod and Hamilton [5] used computer simulations to evaluate the performance of a group of strategies, seeking those that were evolutionarily stable (ESSs). Moreover, they considered not only the final stability of a given strategy, but also the feasibility of each one in an environment dominated by selfish strategies. They found that if the probability of meeting a given partner was above a threshold, then besides the success of ALLD (defect always), another strategy, TFT (tit for tat) emerges as a successful one.

In the tournament organized by Axelrod, 15 strategies were confronted. The winner was TFT. Then the tournament was modified to allow a simulation of evolution [6]. Once more, the winner was TFT. Stemming from Axelrod's article, a great amount of research appeared with interesting modifications in the game. One of them was the

possibility of making wrong moves [10], since it is more realistic to think that animals have random variation in their behavior. In that case the conclusion about the stability of previously analyzed strategies changes remarkably. For example, if two players both play TFT but one of them changes its move from C to D, the opponent will copy that mistake. They will alternate between CD and DC, reducing their performance significantly.

The IPD has been the most used framework for studying the potential of cooperation when there is a short-term temptation to cheat. Despite predictions, it has been very difficult to observe sustained reciprocity in animal cooperation experiments [12]. Recently, Stephens et al. [61] proposed that one possible explanation for the fragility of cooperation in the IPD is strong temporal discounting. This hypothesis is supported by psychological studies showing the animals' preference for receiving small immediate rewards instead of delayed large ones [2, 35, 36, 42].

Although cooperation in the IPD has been studied for over twenty years, most of this research has been focused on strategies that involve nonlearned behavior. Macy and Flache [34] pointed out that game theorists tend to look for solutions for games played by people like themselves. However, the choices that should be made according to the theory have little resemblance to actual decision making. Macy and Flache proposed the use of agent-based models to study the dynamics by which a population moves from one equilibrium to another. They pointed out that, in contrast with the conventional assumption of forward-looking calculation, these models explore backward-looking alternatives based on evolutionary adaptation and learning. Analytical game theory assumes that players have sufficient cognitive skill to make accurate predictions about the consequences of the different decisions; learning theory lightens this requirement by allowing players to base their predictions on their past experience rather than on logical deduction.

Stephens and Clements [60] pointed out that it would be surprising if the mechanisms that enforce animals to achieve a strategic equilibrium were solely genetic. They suggested that a single behavioral mechanism guided by economic principles might provide a mechanism of equilibrium maintenance. Thus, players' selection of the preferred reply might be enforced in the same way that feeding animals track the best way to feed in changing nonstationary environments. Their proposal for learning guided by economic forces was to implement a simple model of Thorndike's law of effect.

Sandholm and Crites [45] evaluated how a popular reinforcement learning algorithm, Q-Learning [68], could learn to play the IPD. The results were discouraging in that Q-Learning could not learn to cooperate with another Q-Learning adversary.

Arita and Suzuki [1] analyzed the interaction between learning and evolution by incorporating the Baldwin [7] effect in the IPD. They have shown that the strategy generated by these mechanisms is an ESS. This work gives insights into how learning can alter the course of evolution in dynamic systems, although neither the mechanism of learning nor the evolution mechanisms used in [1] has biological support. Thus, one interesting proposal is to develop models of learning and evolution with biologically plausible hypotheses in order to analyze their interaction.

Although there are many examples of cooperation by reciprocity in animals [17], the learning mechanisms in cooperation have not been sufficiently analyzed. Since reciprocal altruism is found in different animal species whose evolutionary paths greatly diverge [11, 17], it can be proposed that the mechanisms involved in the learning of cooperation had similar components. Therefore, simple learning mechanisms that are found in a great variety of species and that could be involved in reciprocal altruism should be studied. Many different species are able to change their behavior in order to get larger amounts of appetitive stimuli (food, sex, etc.); this kind of learning—behavior guided by its consequences—is called *operant conditioning*. Operant conditioning

seems to have appeared early in evolutionary history, and it occurs in organisms whose evolutionary paths diverge considerably from that of vertebrates, such as insects like ants [48] and honeybees [22]. These invertebrates have nervous systems that are very different from those of vertebrates; thus, it can be suggested that the ability to learn through operant conditioning has evolved independently in different genetic lines. Here, we propose that the learning of cooperation could have appeared in animals as a mechanism for guaranteeing a higher proportion of appetitive stimuli to them and therefore as a way to increase their own fitness.

In addition to proposing that operant learning can lead to success in the PD against some known strategies, the purpose of this work is to explore the possibility that reciprocal altruism can be explained by simple reinforced learning mechanisms. In this respect, this work is motivated by the finding of Stephens et al. [61] that blue jays learn to play a modified version of the IPD instead of showing innate behavior.

In this article, the role of operant conditioning in the IPD and the EPD will be analyzed. The work is focused on learning mechanisms through a theoretical model presented in [32] and [71] (for both a learning and an evolutionary approach see Arita and Suzuki [1]). That theory (see next sections) provides a new way to understand the role of reinforcement in the learning of cooperation.

2 General Concepts of Operant Conditioning

Staddon [56] pointed out that "Organisms are machines designed by their evolution to play a certain role. This role, together with the environment within which it is played, is called the organism's niche." For simple niches, all that an organism has to do is to make direct responses to specific kinds of stimuli (e.g., most nonsocial invertebrates). The animal does not need to keep any record of its past history to succeed; it is sufficient to avoid aversive stimuli and approach appetitive ones. However, in more complex niches, adaptive behavior requires greater dependence on the animal's past experience that might affect its future behavior in a variety of ways.

The animal's niche affects what it learns and the way that it learns. Niches differ in many respects, and so do learning mechanisms through evolution. However, different niches share many similarities. In the same way, different learning mechanisms have many properties in common. Space and time are common to all niches. The properties of causality—whether an important event is dependent on or independent of a prior event or the animal's own behavior—are also almost the same for every niche. Consequently, a wide range of animal species are able to show adaptation to the temporal, spatial, and causal properties of the environment in similar ways. Like other properties of an organism, the capacity to learn is a product of evolution. Learning occurs because it promotes the propagation of the genetic code of the organism that possesses the capacity to learn [40].

Thus, psychologists have identified classical and operant conditioning as two primary forms of learning that enable animals to acquire relevant characteristics of their environment in order to get reinforcements or to avoid punishments. It is usual to understand classical conditioning as an open-loop experimental procedure where the controlled stimulus delivered by the experimenter is not contingent on the animal's behavior. The learning occurs by repeated association of a conditioned stimulus (*CS*) with an unconditioned stimulus (*US*) that elicits an unconditioned response (*UCR*). For example, in Pavlov's experiment, the dog hears a bell (the *CS*), and after a short time a piece of meat is presented (*US*), which elicits salivation (*UCR*). After repeating the experiment several times, the presentation of the *CS* elicits the salivation response (conditioned response, or *CR*).

On the other hand, operant conditioning is a closed-loop experimental procedure, in the sense that stimuli received by the animal are contingent on its behavior. The animal learns to perform the actions that lead to a reward more frequently, and the ones that lead to punishments less frequently. For example, a pigeon can be trained to press a key when it sees a red light as *CS*, in order to receive a food reward (*US*). The evolutionary advantage of operant conditioning is most clear when the animal's environment is a changing one. A location that once provided food may no longer do so. An unfamiliar animal may turn out to be harmless or a dangerous predator. To be able to survive and pass on its genetic material, an animal must adapt to this variety of situations, and having learning capacities is a fundamental characteristic that permits it to do so.

There are two basically different types of *US*, pleasant and unpleasant ones. Thorndike defined a pleasant stimulus as one that the organism seeks to attain and preserve, and an unpleasant stimulus as one that the organism seeks to avoid or terminate. An experiment can consist in presenting or removing the stimulus in response to a specific behavior. If a pleasant stimulus is presented when the animal performs the desired behavior, the learning is called appetitive. On the other hand, if an unpleasant stimulus is presented and is removed when the animal performs the desired behavior, the learning is called aversive. A brief introduction to appetitive and aversive stimuli is provided in Appendices A and B, respectively.

3 A Theory of Operant Conditioning

3.1 Psychological, Anatomical, and Neurobiological Bases

The theory of adaptive systems, cybernetics, and experimental psychology contributed to the theoretical study of higher brain functions [4, 16, 29, 46, 55, 69]. More recently, fundamental aspects of animal behavior have been included in this approach [46, 47, 55, 58, 59, 71]. From this point of view, Zanutto and Lew [32, 71] presented a neural network model (we will call it ZL) of operant conditioning for appetitive and aversive stimuli.

There are many nonmathematical theories to explain operant conditioning; sometimes there is no agreement about their hypotheses and about the role of prediction, especially in theories to explain escape and avoidance. In the one-factor theory [27], the avoidance response is reinforced by the *US* (e.g., shock). In the two-factor theory [37], instead, the avoidance response is reinforced by the reduction in fear due to the lack of the fear-eliciting *CS*. In the cognitive theory of avoidance [51], it is assumed that during the acquisition phase, animals develop expectations depending on their responses. However, Seligman and his colleagues [39] have proposed that under certain circumstances, animals develop the expectation that their behavior will have little or no effect on their environment, and that this expectation may generalize to a wide range of situations. They called this effect *learned helplessness*. There are also other nonquantitative theories to explain the appetitive data, and only a few mathematical theories to explain operant conditioning. Only a few of them are able to describe the most relevant experimental features for appetitive stimuli [16] and for the aversive stimuli [46, 71].

The main hypotheses of the model are based on psychological, anatomical, and neurobiological bases:

- Behavioral experiments suggest that learning is driven by changes in the expectation about salient future events, mainly reward and punishment. In operant and classical conditioning, the conditioned stimulus (*CS*) anticipates the unconditioned stimulus (*US*). Rescorla and Wagner [43] proposed that animals learn

by comparing what they expect in a given situation and what actually happens. As Staddon [56] has pointed out, animals act because the *CS* allows them to elaborate an expectation or prediction of the unconditioned stimulus. Furthermore, there are neural substrates of prediction and reward, such as the involvement of dopamine neurons of the ventral tegmental area (VTA) and substantia nigra (SN), identified with the processing of prediction and reward [50]. Moreover, Waelti et al. [66] found that the firing of these neurons corresponds with the predictive behavior described by the Rescorla-Wagner rule (explained below). Also, when human beings play the IPD, areas linked with reward processing are consistently activated, such as the nucleus accumbens, caudate nucleus, ventromedial frontal and orbitofrontal cortex, and rostral anterior cingulate cortex [44].

- The pathways that connect the different blocks that were proposed as participating in the operant mechanisms have their bases in anatomical studies in vertebrates such as rats and primates. In monkeys, the prefrontal cortex is a region of convergence of five corticocortical pathways originating in the primary somatic, auditory, visual, olfactory, and gustatory areas. These pathways are relatively independent of one another until they reach the prefrontal cortex, an associative area [19]. In the primate, the prefrontal cortex is the origin of a cascade of reciprocal connective links that flow down from it to the premotor cortex, and from there to the primary motor cortex [8, 67]. As was said before, the VTA is involved in the prediction of the *US*; here its effect is computed by one neuron. The VTA neurons are connected with the prefrontal cortex through the mesocortical dopaminergic system [41].
- In primates, the orbitofrontal cortex is involved in working memory [8, 23]. Neurons in this area continue to discharge for several seconds after the stimulus offset. This short-time memory is able to maintain active useful information to perform specific tasks and to associate *CS*s, actions, and *US*s.
- The cortical effect of the lateral interaction in the premotor and primary motor cortex is simulated by assuming that there are groups of cells that fire in a similar fashion and with high correlation. In this model it is assumed that each response is generated by a cluster of neurons simulated by one neuron representing the effect of the cascade of motor links.
- Learning in this model is controlled by the VTA and SN neurons, which are represented by a prediction neuron. Schultz and colleagues [50] found that there are neurons in the VTA and SN that report an error in the prediction of reward, and when the association between *CS* and *US* is learned, the same neurons fire on the occurrence of the *CS*, predicting that the *US* will follow it. Schultz [49] suggested that the action of a teaching signal can be formalized by applying the Rescorla-Wagner learning rule to synaptic weight. A dopamine teaching signal could modify the weights of the prefrontal synapses according to a Hebbian learning rule. In the proposed model, dopamine neurons control the learning of the prefrontal neurons, by Hebbian or anti-Hebbian learning according to whether the prediction is above a predetermined threshold or not.

In [32] it was shown that the model predicts such relevant features of operant conditioning for appetitive stimulus as the matching law [26], response selection [57], the partial reinforcement extinction effect [30], spontaneous recovery [33], and the successive contrast effect [14]. It can also explain learned helplessness and its reversal [3, 39, 52], delay avoidance [9, 31], and the experiments simulated in [46].

This model is able to explain formally the experimental results on escape and avoidance that were previously interpreted under different theories with contradictory hy-

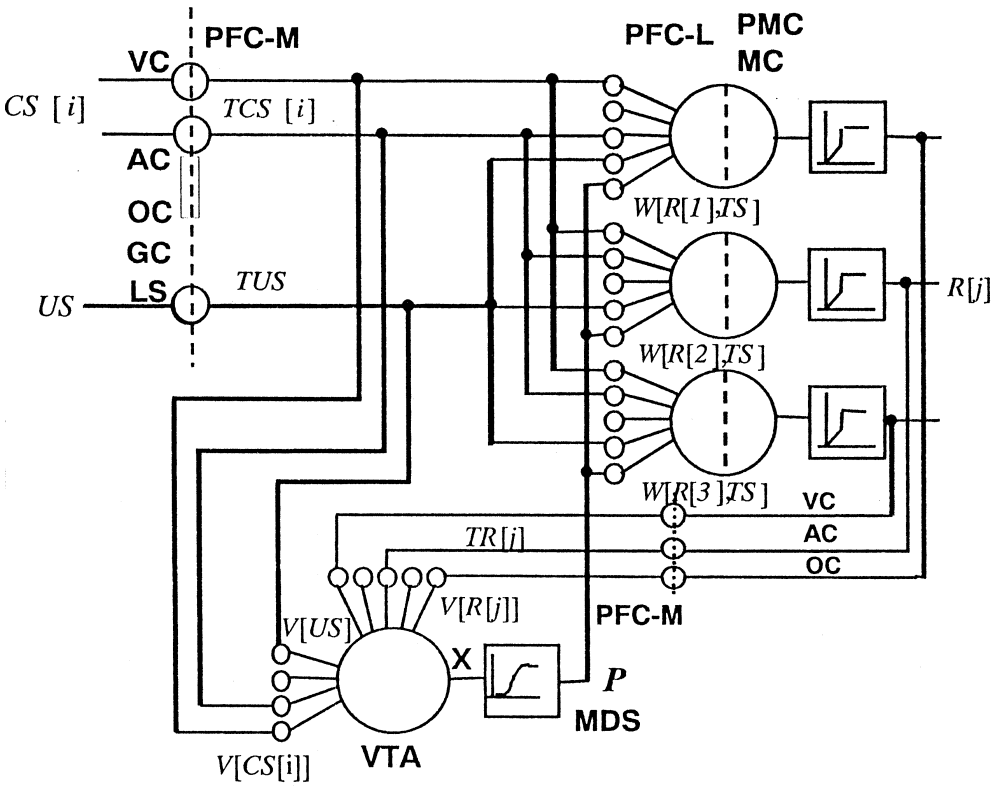

Figure 1. Neural network model. There is one artificial neuron computing the prediction (*P*), and one for each response *R*[*j*] that the animal can make. *P* is a nonlinear function applied to the output of the prediction neuron (*X*). VC: visual cortex; AC: auditory cortex; OC: olfactory cortex; GC: gustatory cortex; LS: limbic system; PFC: prefrontal cortex; PMC: premotor cortex; MC: primary motor cortex; MDS: mesocortical dopaminergic system; VTA: ventral tegmental area. The traces (*TCS*[*i*]) representing the short-term memory of the conditioned stimuli are computed by a group of cells in the PFC-M [23]. The unconditioned stimulus (*TUS*) and the responses (*TR*[*j*]) are inputs to the neurons computing *P* and *R*[*j*]. The synaptic weights *V*[*CS*[*i*]], *V*[*R*[*j*]], and *V*[*US*] represent the associations between the inputs and *P*. The synaptic weights *W*[*R*[*j*], *TS*] are associations between *P*, *TCS*[*i*] and *TUS*, and the PFC-L [8]. The responses performed by the animal are inputs to the different sensory cortices (i.e., VC, AC). These responses generate short-term memory in the PFC-M (bottom of the figure). Finally, the traces from response neurons are fed back to the prediction neuron.

potheses [27, 37, 51]. In [71] it was shown how the model explains experiments that support the two-factor theory. The model also explained how the experiment of Herrnstein and Hineline [28] that supports the one-factor theory does not need a variable such as fear or prediction. The model also explained experiments that support the cognitive theory and show how, under certain circumstances, animals develop the expectation that their behavior will have little effect on their environment (learned helplessness). Finally, this model also explained imitation in the same terms as Schmajuk and Zanutto [46]. In this way, based on biologically plausible hypotheses, this theory explains many relevant features of operant conditioning for appetitive and aversive stimuli.

3.2 The Operant Conditioning Model

The model is shown in Figure 1. The inputs to the model are all the conditioned stimuli (*CS*s) and the unconditioned stimulus (*US*), and the outputs are all the possible animal responses (*R*s). The network has three functional blocks: the stimulus and response traces, the prediction neuron, and the response neurons. In the model, all

the neurons make a summation of their inputs weighted by their synaptic weights, and then a nonlinear function is applied, as is usual in neural network models. Also, synaptic weights are limited in their maximum strength by biological constraints.

The prediction neuron has all the traces of stimuli and responses as inputs. The synaptic weights are modified by the Rescorla-Wagner rule, except for the *US*'s weight, which remains fixed. From the point of view of the two-factor theory, this means that the *US* provokes fear; for the appetitive stimulus, it means that the animal raises its expectation for food (previous models included this hypothesis [46]). The response neurons have all the *CS* and *US* traces and the prediction as inputs. If it exceeds a certain threshold, the learning in the response neurons will be Hebbian in the appetitive, and anti-Hebbian in the aversive, case. If the prediction is below the threshold, the learning is computed inversely. When one of the response neurons exceeds a certain level, the associated response is executed.

3.2.1 Stimulus Traces

There are two types of traces, one corresponding to stimuli, and the other to the responses representing the short-term memory. In the case of the *US*, there are three different traces: short, medium, and long duration. Short-term memories allow the organism to associate a *CS* with a *US* when their presentations are not simultaneous. This hypothesis is supported by neurobiological evidence [23]. The medium- and long-term memories are necessary to explain the influence of reinforcers in learning across several trials. It was previously suggested [16] that medium- and long-time traces can explain the experimental data obtained in learning paradigms such as the partial reinforcement extinction effect (Crespi effect), behavioral contrast, and spontaneous recovery.

The stimulus traces receive as input all the stimuli coming from the visual cortex (VC), auditory cortex (AC), olfactory cortex (OC), gustatory cortex (GC), or limbic system (LS). The outputs of the short-term traces are inputs to response neurons and to the prediction, and the traces from response neurons are fed back to the prediction neuron.

The equation to calculate the short-term traces (T_S) of the stimuli (S) of the *CS* as well as the *US*, at instant n, is a first-order linear difference equation:

$$T_S(n) = T_S(n-1) \cdot (1-\varepsilon) + \varepsilon \cdot S(n) \quad \text{if } S(n) > 0. \tag{3}$$

$$T_S(n) = T_S(n-1) \cdot (1-\beta) \quad \text{if } S(n) = 0 \tag{4}$$

$$T_R(n) = T_R(n-1) \cdot (1-\beta) + \varepsilon \cdot (1 - T_R(n-1)) \cdot R(n) \tag{5}$$

The medium- and long-term traces of the *US* are

$$T_{US\,\mathrm{med}}(n) = T_{US\,\mathrm{med}}(n-1) \cdot (1-\rho) + \rho \cdot US(n) \tag{6}$$

$$T_{US\,\mathrm{long}}(n) = T_{US\,\mathrm{long}}(n-1) \cdot (1-\delta) + \delta \cdot US(n) \tag{7}$$

3.2.2 Prediction Neuron

3.2.2.1 The Rescorla-Wagner Rule

It can be assumed that it is useful for an animal to be able to predict or anticipate important events in its environment, both favorable and unfavorable. Classical conditioning can be seen as a means of learning what *CS*s predict relevant events (*US*s).

The Rescorla-Wagner model states that each time a particular *CS* is presented, one of three things can happen. The first is that the magnitude of the *US* received is higher than expected; in this case, the *CS*s presented shortly before the *US* will strengthen the

UCR. The second possibility is that the *US* received is lower than the expected; in this case, the *CS*s will weaken the *UCR*. The last possibility is that the *US* received is equal to the expected, and no modification occurs. Rescorla and Wagner suggested that the amount of associative strength a *US* center will ultimately support depends on the size of the *US*. They proposed that the change in association (ΔV_i) between CS_i and US_j is given by

$$\Delta V_i = S_i \cdot \left(A_j - \sum_{j=1}^{N} V_i \right)$$

where A_j is the asymptote of associative strength for a given US_j, and V_i is the associative strength between stimulus i and US_j at a specific time. As A_j represents the magnitude of the perceived event, $V = \sum_{i=1}^{n} V_i$ represents what the animal expects to receive in a given situation.

If $A_j > V$, then each time a *CS* is presented in the trial, there will be an increase in the associative strength; if $A_j < V$, there will be a decrease; and if $A_j = V$, there will be no modification. The model states that the learning is proportional to the difference $A_j - V$, with a proportionality factor S_i that represents the salience of the stimulus CS_i. One stimulus can be more salient than another because it is more intense or because it is more noticeable. This simple model is able to explain experimental results such as blocking, extinction, conditioned inhibition, and the overexpectation effect, but it cannot explain experimental results such as latent inhibition.

As stated above, Waelti et al. [66] found that single neurons in the VTA and SN fire according to this formalism. However, the Rescorla-Wagner rule cannot account for time-dependent conditioning phenomena, because it is a trial-based rule. That rule was modified in this operant learning theory to explain real-time conditioning.

3.2.2.2 Model of the Prediction Neuron

The inputs to the prediction neuron are all the short-term traces of *CS*s, *US*, and *R*s. Each response neuron has as input the output of the prediction neuron (P). It has the additional function of controlling its learning. We have

$$X(n) = V_{US}(n)T_{US}(n) + \sum_{i=1}^{N_{CS}} V_{CSi}(n)T_{CSi}(n) + \sum_{i=1}^{N_R} V_{Ri}(n)T_{Ri}(n) \tag{8}$$

$$P(n) = \frac{\xi}{1 + e^{-\upsilon(X(n)-\sigma)}} \tag{9}$$

Here P is the output function, the V's are the weights, and the T's the corresponding traces to the *US, CS*, and *R*. The number of *CS*s is N_{CS}, and the number of responses N_R. The synaptic weight $V_{US}(n)$ remains fixed at 0.1.

If an animal is trained to respond with an appetitive amount of reward, when the amount diminishes, the expectation is modified in such a way that if the reinforcer is still appetitive, animals can stop responding [17]. This effect (some times called the Crespi effect) is simulated by modifying the Rescorla-Wagner model, considering not only the reinforcer value, but also the changes in it. This was done by adding to the reinforcer value the difference between what the animal received in the medium term and in the long term, modulated by a sigmoid. The modulation is to prevent the difference from affecting the intratrial learning, since the Crespi effect is an intertrial effect. This means that if the animal receives a reinforcer of lower value, the sum of the two terms is lower

than the value of the actual reinforcer. The updating of the prediction neuron's weight is based on the Rescorla-Wagner model [43], in which the term $f(DUS(n))$ is added to explain the Crespi effect [14]:

$$VX_S(n) = VX_S(n-1) + \eta(US) \cdot T_S(n) \cdot (US(n) + f(DUS(n)) - X(n)) \tag{10}$$

$$V_S(n) = \frac{2}{1 + e^{-\kappa \cdot VX_S(n)}} - 1 \tag{11}$$

The associative strength is represented by $VX_S(n)$. Equation 11 clamps the synaptic weights in the range of -1 to 1. The VX values are bounded between 10 and -10 in order to limit the maximum associative strength. Here the salience is represented by the stimulus trace $T_S(n)$; this means that a *CS*'s salience depends on its memory trace. The rate of learning is represented by $\eta(US)$, which depends on whether the *US* is present or not, due to attentional modulation:

$$\eta(US) = \eta_i \quad \text{if} \quad US > 0, \quad \text{and} \quad \eta(US) = \eta_d \quad \text{if} \quad US = 0$$

The term $f(DUS(n))$ is defined in the following way:

$$DUS(n) = T_{US\,\mathrm{med}}(n) - T_{US\,\mathrm{long}}(n) \tag{12}$$

$$f(DUS(n)) = \chi \cdot (\tau + DUS(n)) \cdot \tanh(\gamma \cdot DUS(n))^{10} \tag{13}$$

3.2.3 Response Neurons

As was said above, there is an output neuron for each of the possible responses of the animal. The output of these neurons is determined by

$$R_j(n) = g(Y_j(n))$$

$$Y_j(n) = W_{j\,\mathrm{pred}}(n) \cdot P(n) + W_{j\,US}(n) \cdot T_{US}(n) + \sum_{i=1}^{N_{CS}} W_{j\,CSi}(n) \cdot T_{CSi}(n) + noise(n)$$

$$g = 0 \quad \text{if } Y_j(n) < 0; \qquad g = 1 \quad \text{if } Y_j(n) > \mu; \qquad \text{else} \quad g = Y_j(n) \tag{14}$$

where $W_{j\,\mathrm{pred}}(n)$ is the jth response synaptic weight corresponding to the output of the prediction neuron at instant n, $W_{j\,US}(n)$ is the weight corresponding to the *US*, and $W_{j\,CSi}(n)$ is the weight corresponding to each of the *CS*s. The output of the prediction neuron is $P(n)$, the *US* short-term memory is $T_{US}(n)$, and the *CS* short-term memories are $T_{CSi}(n)$.

The animal executes a response R_j whenever Y_j exceeds the threshold μ. At any instant, only one response can be executed. The updating is done asymmetrically. At each instant, one neuron is selected randomly, and only its weights are updated.

If no response is executed after a fixed time, a neuron is randomly selected. This differs from [32], where it is possible that during a trial no response is executed. However, in the particular case of the PD it is mandatory to choose a determined move.

The equation to compute the learning of these neurons is based on the Hebb rule [25], which states that the change in synaptic efficacy is proportional to the presynaptic and postsynaptic activity. If the *US* is predicted, the learning will be Hebbian in the appetitive case. The association between stimuli and the selected response will be reinforced because it produces the procurement of a positive reinforcement. The weights

of the executed neuron are updated in the following way:

$$W_{jq}(n) = \psi W_{jq}(n-1) + \phi\Omega Q(n) T_{Rj}(n) \tag{15}$$

where Q is the input (P, T_{US}, or T_{CSi}), q is the respective index (P, US, or i), and $T_{Rj}(n-1)$ is the jth short-term response trace at instant $n-1$. The coefficient ψ is a constant that represents the synaptic weight forgetting rate. The learning rate is controlled by ϕ and Ω. Here Ω can take either of two values: if the *US* is appetitive and $P < \lambda$, then $\Omega = -\lambda$, and if $P \geq \lambda$ then $\Omega = \lambda$. The reverse rule is applied in the case that the *US* is aversive. The constant λ is the learning threshold: if the prediction is higher, it means that the active *CS*s will signal that a *US* is likely to come; if the *US* is appetitive, the response that leads to reinforcement procurement will be strengthened.

3.3 How the Model Works

In order to understand the basic workings of the model, we will explain how the animal can associate the presentation of a stimulus, a selected action, and an appetitive reinforcement. Let us suppose that a light is turned on for 5 seconds (*CS*), and if the animal presses a certain lever, it receives food (*US*). The light, coming from the visual cortex, generates a short-term memory. If the animal executes the appropriate action by chance, it will receive a reinforcement. Since the synaptic weight of the *US* in the prediction neuron is 0.1, if the *US* is sufficiently high it will make the prediction neuron fire with an output above the threshold, raising the synaptic weight of the *CS* connected to the right response neuron and the one corresponding to the prediction. In addition, as there is a discrepancy between the received *US* and what the prediction neuron has predicted, the synaptic weight of the *CS* connected to the prediction neuron will be updated in proportion to the discrepancy between the *US* and the prediction. In the following trials, the correct answer will have a higher chance than the others to be selected, since the *CS* will have a synaptic weight greater than 0. The *CS*'s synaptic weight in the prediction neuron will rise gradually until it can completely predict the *US*. The sole presentation of the *CS* will make the prediction neuron fire, and the animal will respond sooner to the presentation of the light. The reverse happens when the animal chooses an incorrect answer: The prediction neuron does not fire, the learning becomes anti-Hebbian, and the synaptic weight that associates the *CS* with the incorrect response neurons decreases, causing its probability of firing to be reduced in subsequent experiments. It is important to notice that without short-term memory these associations could not be made, since the *CS* offset is previous to the *US* presentation.

4 Iterated Prisoner's Dilemma

In the present experiment, we evaluated the role of operant conditioning in the learning of cooperation in the IPD game. We compared its performance against some previously commonly proposed strategies. Each pair of players was matched in 1000 rounds. We performed 100 repetitions to calculate the average. The same analysis was done with four different probabilities of making a random move. The chosen probabilities (*pr*) were 0, 0.01, 0.02, and 0.05. In the case where two ZL players are matched, two independent implementations of the ZL model that learn separately from each other are used.

The inputs to the model are a stimulus to signal the beginning of a new trial, where the animal chooses to cooperate or to defect, and two more stimuli; one indicates that the opponent's move was C in the previous round, and the other indicates that it was D. The response is computed in 100 time units. The two participants received points according to Table 1, depending on the actions performed by both players; here the

US value is 2 times the number of earned points (to keep the same constants as in the ZL model).

One critical issue for an operant behavior experiment is to decide whether the *US* is rewarding (appetitive) or punishing (aversive). The operant learning model used here is able to learn from both appetitive and aversive stimuli. However, during an experiment it has to receive all appetitive or all aversive reinforcers; it cannot alternate between rewarding and punishing. Here, the model was used in its appetitive alternative; this means that even if it does not receive any reward, the model will not consider that it was punished. If the opponent decides to defect, the model will receive a low reward that will not be enough to reinforce any response. In the terms of the theory, due to the low value of each reinforcer when the opponent defects, the output of the prediction neuron is below the Hebbian learning threshold, and no response is reinforced. When animals learn that their behavior has little effect on their environment, they stop responding (this is called learned helplessness).

The model is able to predict this behavior. Let us suppose that an animal can press one lever that corresponds to the action of cooperation, or another that corresponds to the action of defection. If the animal is playing against ALLD, receiving very little food, it will press each lever randomly, and after several trials it will stop responding. Thus, the animal will not learn to defect against an ALLD strategy, and in an IPD game it will be exploited by it. This means that operant conditioning alone is not able to survive in such conditions. We propose that in addition to operant conditioning, in order to be able to learn to play the IPD, a mechanism of protection against defectors is needed; otherwise, cooperation will be an advantage to the selfish player, and animals that cooperated according to an operant behavior would be exploited by the defectors. This assumption poses an additional hypothesis to be tested in animals that cooperate according to reciprocal altruism.

We simulated this requirement by a long-time memory trace of the opponent's defections. If the level of the trace (TR) is above a certain defection threshold $th = 0.9$, then ZL will execute D ($\alpha = 0.05$):

$$TR_D(n) = \alpha \cdot D(n) + (1 - \alpha) \cdot TR_D(n - 1) \tag{16}$$

Table 2 summarizes the strategies used in the simulations. To evaluate different experiments, two groups of strategies are analyzed. Group 1 comprises the strategies that the behavior of which is not influenced by the actions of the adversary. They are ALLD, ALLC, ALTDC, RANDOM. Group 2 comprises the strategies that change their actions depending on the behavior of the opponent. They are ZL, TFT, PAVLOV, S_MAJO, TF2T, and S_TFT.

4.1 IPD Results

Figures 2 and 3 show ZL's average score against group 2 adversaries for $pr = 0$ and 0.02 for iteration. To compare the general performance of each of the strategies against the others, confrontations were simulated between each pair. Tables 3–6 show the results for different probabilities of random moves: 0, 0.01, 0.02, and 0.05.

When ALLD played against ZL, the first strategy got a higher score than the second one, because it takes some rounds to exceed the defection threshold. ZL did not cooperate in all the encounters with ALLC, as group 2 strategies do. This is because whatever the ZL's response is, it obtains enough reinforcement.

Because ZL has to learn a strategy in order to obtain reinforcers, it takes more time to stabilize the responses. Consequently, the average score obtained by ZL against another ZL was a little lower than against TFT. This was a disadvantage in the EPD (see below). Figure 4 shows the total score for each strategy and the probability of wrong

Table 2. Description of the strategies employed.

Strategy	Description
TFT	First move C, and then repeat the adversary's previous move.
ALLC	Always C.
ALLD	Always D.
ALTDC	Alternate D and C.
ZL	Move depends on the response of the model's neurons.
RANDOM	D with probability 0.5, C otherwise.
SOFT MAJORITY (S_MAJO)	C on the first move, and then play the opponent's most used move (C in case of equality).
SLOW TFT (S_TFT)	C on the first two moves, then begin to defect after two consecutive D of its opponent, and return to cooperation after two consecutive C.
PAVLOV	C on the first move, and then cooperate only if the two players made the same move.
TF2T	C on the first move; then D if the opponent has defected two consecutive times, and C otherwise.

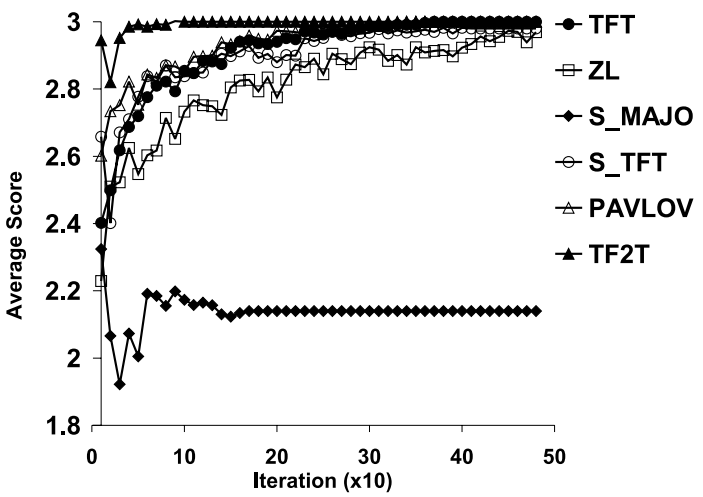

Figure 2. Average score obtained by ZL against group 2 opponents as a function of the iteration number ($pr = 0$). ZL learned faster against genetic strategies than against another ZL. The only scheme that ZL did not learn to cooperate with was S_MAJO.

moves. ZL got the second highest total score with $pr = 0$ (after TFT), and for the other prs, ZL was the winner. Figure 5 shows the score of ZL against each strategy with the different pr.

The cooperation level decreases with higher pr. However, with $pr = 0.05$, ZL and TFT cooperated 75% of the time, and ZL against itself cooperated 66% of the time.

5 Evolutionary Prisoner's Dilemma

Tables 3–6 present the average score for each pair of strategies in a 1000-round game. In the actual experiment, those scores were taken to simulate an evolution of the

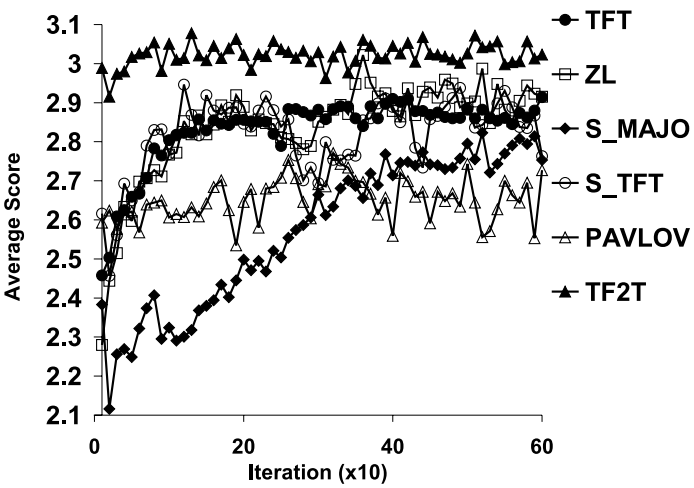

Figure 3. Average score obtained by ZL against group 2 opponents as a function of the iteration number ($pr = 0.02$). Even with wrong actions, ZL learned to cooperate. These random moves allow better performance of ZL against S_MAJO.

Table 3. Average score in 1,000 rounds for the strategy in the row label when matched against the strategy in the column label: $pr = 0$.

	TFT	ALLC	ALLD	ALTDC	ZL	RANDOM	S_MAJO	S_TFT	PAVLOV	TF2T	Total
TFT	3.000	3.000	0.999	2.500	**2.957**	2.248	3.000	3.000	3.000	3.000	26.704
ALLC	3.000	3.000	0.000	1.500	**1.440**	1.503	3.000	3.000	3.000	3.000	22.443
ALLD	1.004	5.000	1.000	3.000	**1.089**	2.996	1.004	1.008	3.000	1.008	20.109
ALTDC	2.500	4.000	0.500	2.000	**2.111**	2.254	2.500	4.000	2.250	4.000	26.115
ZL	**2.957**	**4.040**	**0.978**	**2.347**	**2.902**	**2.284**	**2.139**	**2.943**	**2.970**	**2.997**	**26.558**
RANDOM	2.250	3.998	0.501	2.250	**2.177**	2.245	2.289	2.253	2.249	3.126	23.337
S_MAJO	3.000	3.000	0.999	2.500	**2.183**	2.236	3.000	3.000	3.000	3.000	25.918
S_TFT	3.000	3.000	0.998	1.500	**2.950**	2.246	3.000	3.000	3.000	3.000	25.695
PAVLOV	3.000	3.000	0.500	2.250	**2.966**	2.253	3.000	3.000	3.000	3.000	25.969
TF2T	3.000	3.000	0.998	1.500	**2.986**	1.873	3.000	3.000	3.000	3.000	25.356

Table 4. Average score in 1,000 rounds for the strategy in the row label when matched against the strategy in the column label: $pr = 0.01$.

	TFT	ALLC	ALLD	ALTDC	ZL	RANDOM	S_MAJO	S_TFT	PAVLOV	TF2T	Total
TFT	2.300	3.005	1.009	2.492	**2.913**	2.248	3.005	2.812	2.287	3.004	25.072
ALLC	2.980	2.995	0.020	1.508	**1.479**	1.512	2.994	2.995	1.551	2.995	21.027
ALLD	1.038	4.970	1.014	2.992	**1.106**	2.995	1.019	1.023	2.992	1.063	20.211
ALTDC	2.492	3.982	0.518	2.004	**2.201**	2.242	2.333	2.272	2.251	3.966	24.260
ZL	**2.892**	**4.001**	**0.993**	**2.281**	**2.816**	**2.273**	**2.458**	**2.885**	**2.789**	**3.021**	**26.410**
RANDOM	2.249	3.979	0.517	2.253	**2.188**	2.252	2.136	2.252	2.248	3.119	23.195
S_MAJO	2.979	2.995	1.012	2.334	**1.694**	2.300	2.995	2.995	2.428	2.997	24.728
S_TFT	2.792	2.995	1.014	2.239	**2.876**	2.248	2.995	2.937	2.960	2.995	26.049
PAVLOV	2.288	3.953	0.518	2.250	**2.932**	2.253	2.093	1.223	2.976	2.847	23.332
TF2T	2.980	2.995	1.003	1.515	**2.775**	1.874	2.993	2.994	2.091	2.996	24.213

Table 5. Average score in 1,000 rounds for the strategy in the row label when matched against the strategy in the column label: $pr = 0.02$.

	TFT	ALLC	ALLD	ALTDC	ZL	RANDOM	S_MAJO	S_TFT	PAVLOV	TF2T	Total
TFT	2.296	3.009	1.020	2.486	**2.893**	2.251	2.990	2.466	2.275	3.009	24.694
ALLC	2.961	2.989	0.041	1.515	**1.492**	1.518	2.990	2.988	1.495	2.989	20.975
ALLD	1.073	4.939	1.028	2.984	**1.124**	2.987	1.033	1.041	2.986	1.115	20.309
ALTDC	2.486	3.966	0.539	2.011	**2.125**	2.248	2.237	2.300	2.252	3.930	24.092
ZL	**2.852**	**3.981**	**1.005**	**2.329**	**2.717**	**2.276**	**2.667**	**2.823**	**2.641**	**3.026**	**26.317**
RANDOM	2.255	3.964	0.535	2.255	**2.198**	2.245	2.309	2.258	2.251	3.111	23.382
S_MAJO	2.941	2.989	1.029	2.350	**1.648**	2.217	2.990	2.989	2.555	2.991	24.699
S_TFT	2.444	2.991	1.028	2.228	**2.805**	2.243	2.990	2.790	2.926	2.979	25.423
PAVLOV	2.276	3.979	0.535	2.250	**2.872**	2.257	1.644	1.156	2.952	2.838	22.756
TF2T	2.959	2.990	1.008	1.530	**2.658**	1.881	2.988	2.980	2.055	2.989	24.036

Table 6. Average score in 1,000 rounds for the strategy in the row label when matched against the strategy in the column label: $pr = 0.05$.

	TFT	ALLC	ALLD	ALTDC	ZL	RANDOM	S_MAJO	S_TFT	PAVLOV	TF2T	Total
TFT	2.276	3.019	1.046	2.465	**2.792**	2.249	3.023	2.213	2.262	3.018	24.363
ALLC	2.905	2.972	0.098	1.539	**1.544**	1.540	2.975	2.963	1.545	2.972	21.050
ALLD	1.171	4.853	1.073	2.966	**1.188**	2.967	1.081	1.091	2.964	1.265	20.618
ALTDC	2.463	3.912	0.583	2.019	**2.132**	2.252	2.298	2.310	2.252	3.832	24.051
ZL	**2.710**	**3.908**	**1.043**	**2.323**	**2.574**	**2.275**	**2.415**	**2.743**	**2.434**	**3.010**	**25.436**
RANDOM	2.254	3.911	0.586	2.248	**2.197**	2.255	2.268	2.254	2.246	3.083	23.303
S_MAJO	2.900	2.972	1.074	2.281	**1.493**	2.239	2.971	2.965	2.618	2.971	24.482
S_TFT	2.198	2.982	1.072	2.221	**2.714**	2.248	2.982	2.307	2.832	2.918	24.472
PAVLOV	2.259	3.907	0.587	2.248	**2.744**	2.248	1.429	1.297	2.877	2.828	22.422
TF2T	2.902	2.976	1.025	1.572	**2.515**	1.898	2.977	2.909	2.033	2.975	23.780

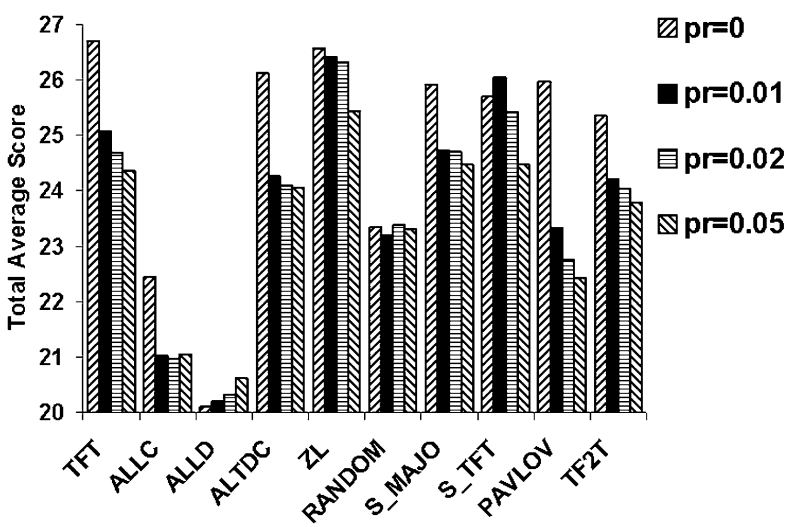

Figure 4. Total average score for each strategy and probability of random moves. TFT obtained the highest score for $pr = 0$, followed closely by ZL. For the other prs, ZL was the best strategy. Its performance was not significantly affected by pr.

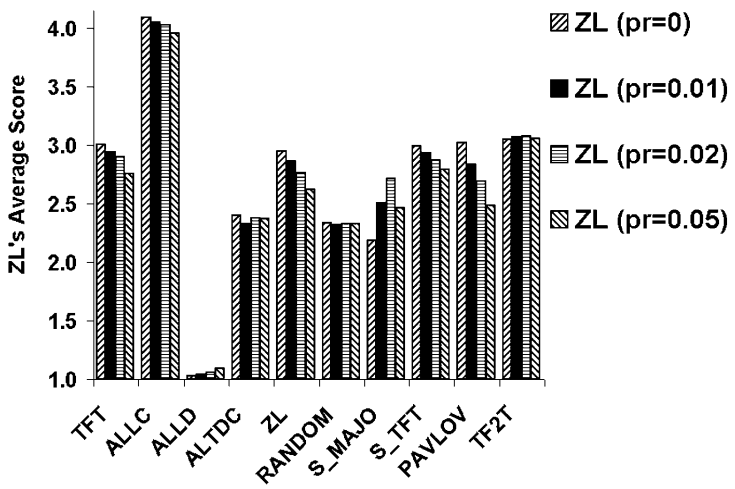

Figure 5. Average score obtained by ZL against each scheme for the different *pr*. The most important decrease in performance was against PAVLOV. There is an important increment against S_MAJO at higher *pr*. The random moves did not significantly change the score obtained in the other cases.

population. The new generation is formed with a different arrangement of players, proportional to the total score obtained against all other adversaries. Thus, the total score for each strategy changes in each new generation, because the proportion of opponents changes.

Let M be the matrix of average scores between each pair of strategies, and m_{ij} the average score that strategy i gets against j. Let $P(n)$ be the column vector that indicates the proportion of each strategy in generation n, and $T(n)$ the total average score in generation n for each player. Then

$$M \cdot P(n) = T(n) \tag{17}$$

The total average score is used to calculate the population of the next generation. For each strategy i,

$$P_i(n) \propto P_i(n-1) \cdot T_i(n-1) \tag{18}$$

We will present results on the evolution of populations in different arrangements. In the first experiment, the initial population consisted of ZL and group 1 strategies. All the strategies of group 1 are degenerate and it is not trivial that ZL can survive in such conditions. The behavior of ZL cannot be predicted easily without simulating it. The second experiment consisted of all group 2 strategies. In the last experiment, all the schemes were evaluated together. Finally, we studied the influence of learning in ZL's performance by pretraining it against all the other strategies and then analyzing its behavior in the EPD simulation.

5.1 EPD Results

In the first experiment, we analyzed strategies with a population from group 1 and ZL. We examined how ZL evolves in a population composed of the static schemes

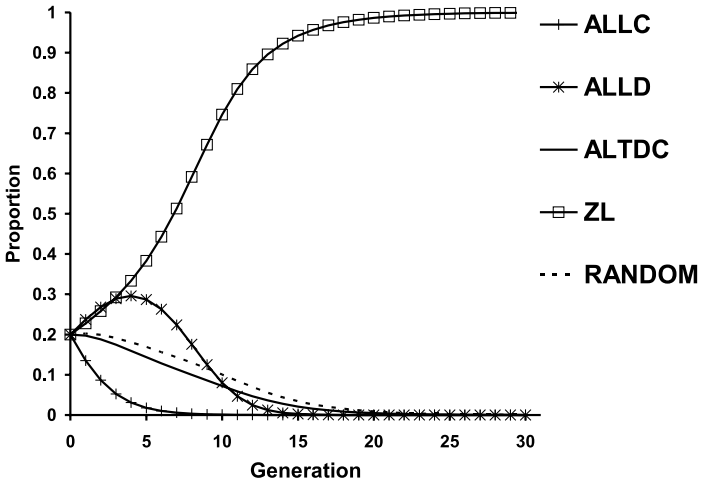

Figure 6. ZL and group 1 schemes in the EPD for $pr = 0$. In few generations, ZL reached 100% of the population.

(ALLD, ALLC, ALTDC, RANDOM). Figure 6 shows that ZL eliminated the others in a few generations. The same effect happened for the other probabilities of random moves.

In the second experiment, we evaluated the evolution of group 2 schemes. The aim was to find if ZL could survive in an environment with players that punish selfish behavior. All the strategies, except ZL, correspond to a genetic trait. This is a disadvantage for ZL, since it needs more time to reach its asymptotic behavior, decreasing its fitness. This disadvantage is more clearly seen when results are compared with experiments where ZL was previously pretrained against all other strategies (see below). It was the only strategy that was extinguished when $pr = 0$. However, when random moves were introduced, ZL survived throughout the generations. The results are shown in Figures 7 to 10.

In the last experiment, all 10 schemes were evaluated, starting from a homogeneous population. In Figure 11, with $pr = 0$, group 2 strategies (except ZL) and ALLC persisted in the population throughout the generations. However, with $pr = 0.01$ (Figure 12) the only two strategies that survived after a series of oscillations were S_MAJO and TFT, the latter being scarce.

Table 7 summarizes the final population of each strategy in the different experiments.

Finally, we show that ZL is not extinguished in any of the previous experiments when it is first pretrained. We present only the EPD results where ZL extinguished previously (group 2 with $pr = 0$, and all the schemes with $pr = 0$ and $pr = 0.01$). Table 8 summarizes the final population of each strategy in the different experiments when ZL is pretrained.

6 Discussion

In the IPD (Figure 2), ZL learned to cooperate with all the strategies of group 2, except with S_MAJO for $pr = 0$. This was because S_MAJO remembers all the history of the opponent actions. For this reason, it defects unless others change their actions. Thus, S_MAJO does not help to reverse a noncooperative tendency of the adversary. If the round is prolonged for a certain time, then, due to S_MAJO's defection in each round,

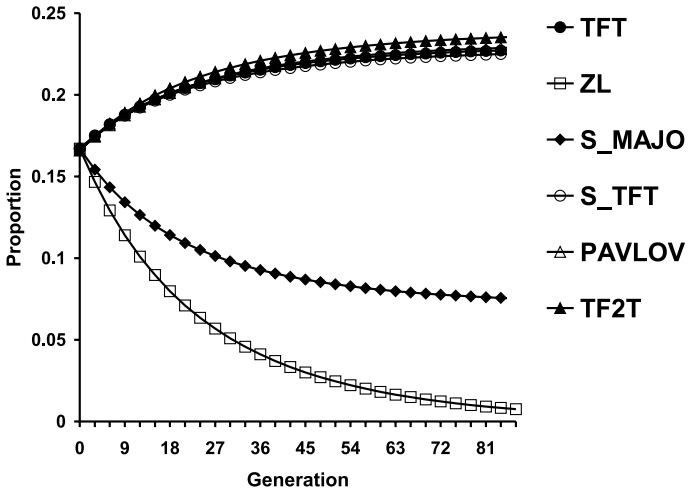

Figure 7. Group 2 schemes in the EPD for $pr = 0$. Despite the fact that ZL learned to cooperate with group 2 strategies, it was extinguished in this case. This was because ZL is a learned strategy, and it needs more time to reach its asymptotic behavior. In this case, ZL had disadvantages against genetic schemes.

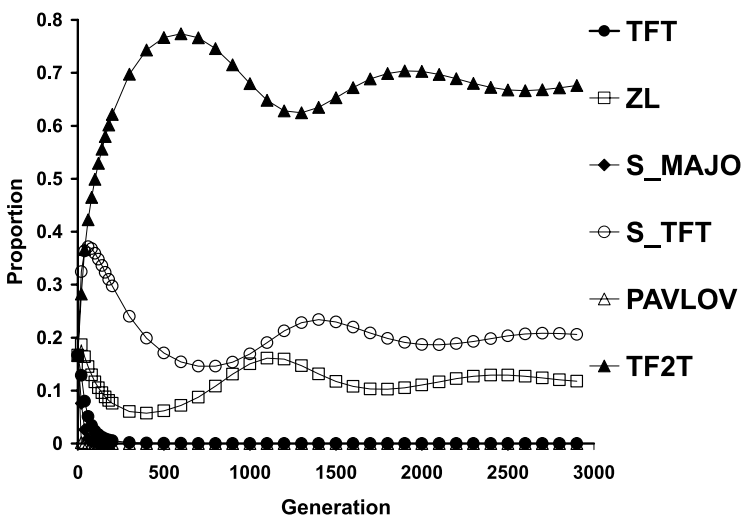

Figure 8. Group 2 schemes in the EPD for $pr = 0.01$. A small probability of random moves caused the extinction of TFT, S_MAJO, and PAVLOV. On the other hand, ZL reached stability in the population, despite its initial disadvantage.

ZL begins to defect all the time because the threshold in Equation 16 is exceeded. After some iterations, any action changes the behavior of S_MAJO very little. From an operant point of view, ZL's actions are not contingent on the obtained reinforcement. However, these strategies learned to cooperate more frequently when $pr = 0.02$ (Figure 3). This was because S_MAJO could change its tendency because some actions were randomly chosen.

We showed that ZL learned faster against TF2T, S_TFT, PAVLOV, and TFT than against another ZL (Figure 2). This was because the correlation between any of their actions and ZL is stronger than with another ZL, due to its exploratory behavior. ZL learned

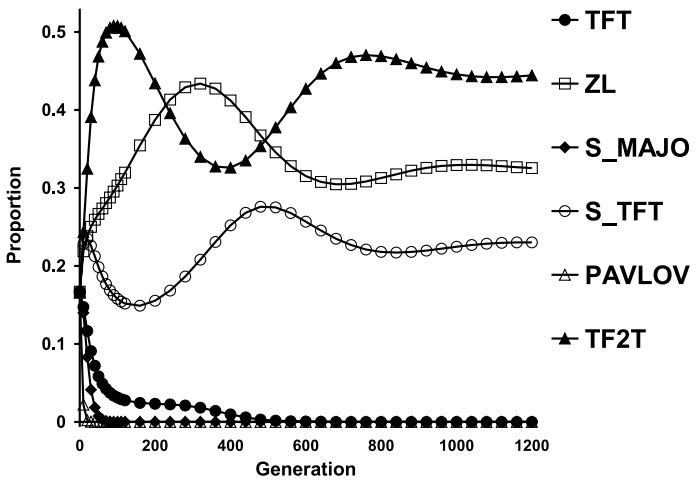

Figure 9. Group 2 schemes in the EPD for $pr = 0.02$. In this case, TFT, TF2T, and ZL composed the final population.

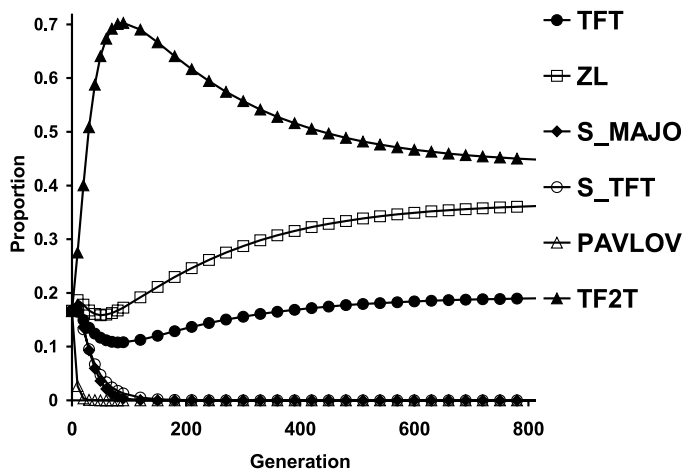

Figure 10. Group 2 schemes in the EPD for $pr = 0.05$. In the final population were TFT, TF2T, and ZL.

to cooperate at a high rate even with $pr = 0.05$ (see Tables 2–5). In spite of the random moves, the average score of ZL against each strategy did not vary significantly, as happened with the others (Figure 5).

In the EPD we wanted to study how ZL evolved in a different population arrangement. When we compared it with strategies of group 1, ZL was the dominant scheme even with wrong moves. If we replace ZL with TFT and $pr = 0.01$ or higher, TFT is no longer the only strategy that survives throughout the generations.

When ZL was compared with group 2 strategies and $pr = 0$, all schemes survived except ZL. The reason is that ZL takes more time to cooperate, because, unlike the others, it has to learn. The only players that survived for $pr = 0.01$ and 0.02 are ZL, S_TFT, and TF2T. For $pr = 0.05$, the survivors were ZL, TF2T, and TFT.

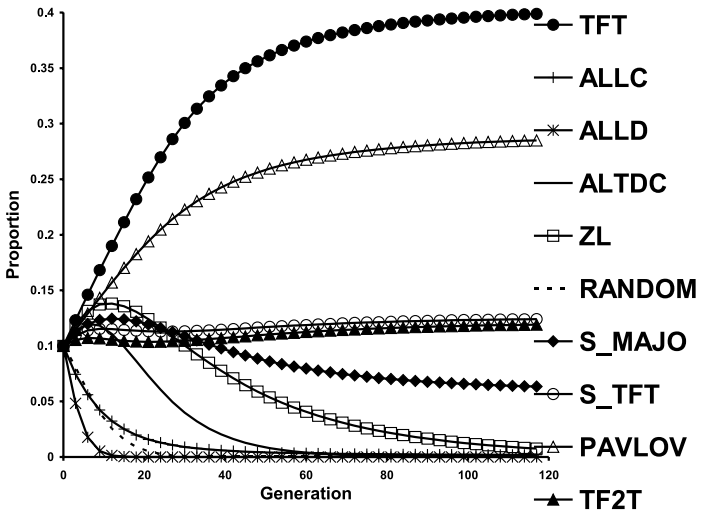

Figure 11. All schemes in the EPD for $pr = 0$. With the exception of ZL, the only strategies that survived were from group 2.

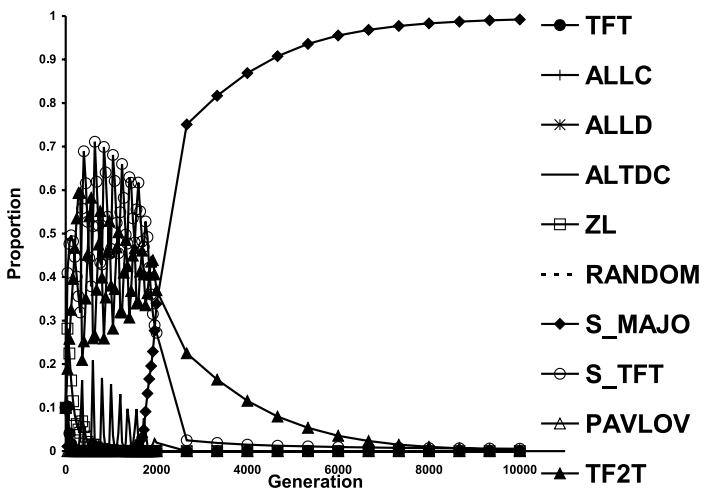

Figure 12. All schemes in the EPD for $pr = 0.01$. After a series of oscillations, S_MAJO raised its proportion from almost 0% to near 100%.

In the last experiment, the evolution of all strategies was studied (Figures 11–14). When $pr = 0.01$, at the beginning S_MAJO and TFT were almost extinguished, and TF2T, S_TFT, and ALTDC were the dominating schemes. S_MAJO increased in proportion when ZL decreased, since S_MAJO gets the least score against ZL, with the exception of ALLD (see Table 4). After a series of oscillations, S_MAJO reaches almost 100% of the population, showing the important role of the participants' interplaying in the population's dynamics.

ZL stopped extinguishing itself at prs of 0.02 and 0.05 when the other survivors were TF2T, S_TFT, and ALTDC. Figure 14 shows that ZL had the greatest proportion at $pr = 0.05$. EPD results showed that ZL survives against group 2 strategies only if the

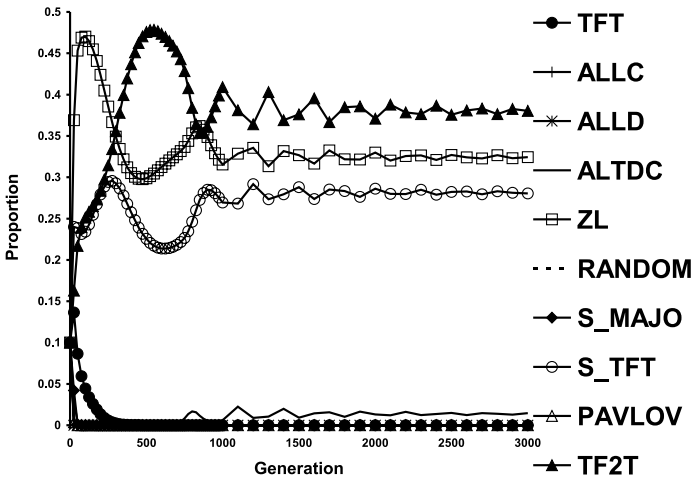

Figure 13. All schemes in the EPD for $pr = 0.02$. The final population is composed of TF2T, ZL, S_TFT, and a small proportion of ALTDC.

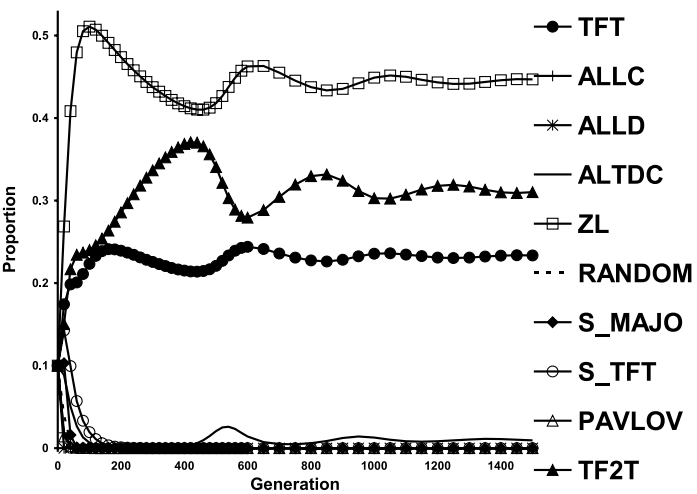

Figure 14. All schemes in the EPD for $pr = 0.05$. ZL had the highest proportion in the final population. The other survivors were TF2T, TFT, and to a small extent ALTDC.

moves are noisy. This is because the other strategies can achieve perfect cooperation, while ZL is always below its ideal score because it loses points while learning the best reply to the opponent. However, when random moves are allowed in the game, ZL showed a relatively low sensitivity to the noise, and the initial advantage that the other strategies have diminished. In this way, ZL is able to survive under these conditions.

Animals learn to respond correctly to a *CS*, even when the *US* is presented with a certain probability. Thus, animals do not extinguish or change their responses when their actions do not produce what they expect for some trials without reward. As stated previously, the results also showed that ZL can overcome its disadvantage against group 2 strategies when there is some probability of making random moves. On the other

Table 7. Final proportion of each strategy in the EPD experiments.

	Proportion											
	All				Group 1				Group 2			
Strategy	$pr = 0$	$pr = 0.01$	$pr = 0.02$	$pr = 0.05$	$pr = 0$	$pr = 0.01$	$pr = 0.02$	$pr = 0.05$	$pr = 0$	$pr = 0.01$	$pr = 0.02$	$pr = 0.05$
TFT	0.404	0.0145	0	0.2325	—	—	—	—	0.23	0	0	0.1931
ALLC	0.002	0	0	0	0	0	0	0	—	—	—	—
ALLD	0	0	0	0	0	0	0	0	—	—	—	—
ALTDC	0	0	0.0138	0.0099	0	0	0	0	—	—	—	—
ZL	0	0	0.3245	0.4449	1	1	1	1	0	0.1201	0.3246	0.3684
RANDOM	0	0	0	0	0	0	0	0	—	—	—	—
S_MAJO	0.06	0.9845	0	0	—	—	—	—	0.072	0	0	0
S_TFT	0.126	0	0.2816	0	—	—	—	—	0.228	0.2009	0.2275	0
PAVLOV	0.289	0	0	0	—	—	—	—	0.232	0	0	0
TS2T	0.121	0	0.3802	0.3127	—	—	—	—	0.239	0.679	0.4479	0.4385

Table 8. Final proportion of each strategy in the EPD experiments when ZL is pretrained against all other strategies.

	Proportion											
	All				Group 1				Group 2			
Strategy	$pr = 0$	$pr = 0.01$	$pr = 0.02$	$pr = 0.05$	$pr = 0$	$pr = 0.01$	$pr = 0.02$	$pr = 0.05$	$pr = 0$	$pr = 0.01$	$pr = 0.02$	$pr = 0.05$
TFT	0.393	0.1115	0.2033	0.2402	—	—	—	—	0.249	0.1115	0.2034	0.2402
ALLC	0	0	0	0	0	0	0	0	—	—	—	—
ALLD	0	0	0	0	0	0	0	0	—	—	—	—
ALTDC	0	0	0	0	0	0	0	0	—	—	—	—
ZL	0.085	0.8885	0.7965	0.4457	1	1	1	1	0.003	0.8885	0.7966	0.4457
RANDOM	0	0	0	0	0	0	0	0	—	—	—	—
S_MAJO	0	0	0	0	—	—	—	—	0.003	0	0	0
S_TFT	0.13	0	0	0	—	—	—	—	0.249	0	0	0
PAVLOV	0.278	0	0.0002	0	—	—	—	—	0.248	0	0	0
TS2T	0.114	0	0	0.3141	—	—	—	—	0.248	0	0	0.3141

hand, the other strategies used in this article change their behavior in a predetermined way. For example, if two players both play TFT but one of them changes its move from C to D, the opponent will copy that mistake, resulting in an alternation between CD and DC. The operant model does not change its behavior in any such dramatic way, preventing the mistaken move from propagating. Other strategies, such as TF2T, S_TFT, PAVLOV, and S_MAJO, show less sensitivity to noise than TFT.

Finally, we pretrained ZL against all the other strategies to demonstrate that ZL loses valuable fitness because of its learning mechanisms (see Figures 15–17). The results showed that ZL survives in all the proposed arrangements, and when there is some probability of random moves, it achieves the major proportion in the population. These results support the idea that although learning has the important property of adapting to a variety of strategies, the time that learning takes represents a disadvantage in evolutionary terms compared with nonlearning strategies. Two mechanisms that ZL can use to overcome this disadvantage are imitation and the Baldwin effect.

The first mechanism provides a ZL player the possibility of observing how another ZL confronts other players by imitating their moves. This allows the first ZL player to learn other strategies without the disadvantage of losing fitness. It has to be mentioned that there is no need to change the model, since it can already learn by imitation [46].

The second mechanism, the Baldwin effect, consists of the fact that organisms that make nonhereditary physical or behavioral modifications to survive environmental dangers will have a higher proportion in future generations than less adaptive animals. Through natural selection, then, the capacity for such adaptation along with the beneficial acquired traits will become universal. Simpson [53] reintroduced the Baldwin effect as genetical reinforcement of advantageous but initially nonhereditary traits. He stated that a genetic version of a seemingly nonhereditary adaptation may arise when natural selection acts on the likelihood of having an adaptive trait not just on the trait itself. Thus we may study the possibility that reciprocal altruism could have started as a learned behavior, and then been incorporated genetically by means of the Baldwin effect.

The experiments showed that our operant learning theory explained how an individual learns to play the PD. However, reciprocal altruism has other properties than the PD. Reciprocal altruism is not limited by simultaneous interactions as in the PD. In addition, it requires that individuals recognize the other participants. These properties can be simulated with an operant model, but that is not the purpose of this article.

The IPD has been the most used framework to study non-kin cooperation. However, it has been shown experimentally that animals show a strong tendency to defect [12, 18, 20, 21]. Stephens et al. [61] proposed that this happens because animals strongly discount future reward. They showed that cooperation between blue jays can be stabilized if the influence of temporal discounting is diminished by implementing payoff accumulation. They conducted a factorial experiment, manipulating discounting and the strategy that blue jays confronted. It was shown that when the birds played against an ALLD strategy, they defected most of the time regardless of whether they accumulated the food or not. But when the birds played against a TFT strategy, the frequency of cooperation was high in the case of food accumulation and was very low in the other case. From this study we conclude that the role of reinforcement in the animal's behavior has many other characteristics and consequences that are not considered in the IPD and that must be taken into account.

With the operant conditioning model, different relations between the payoffs and reinforcement can be studied. Here, the amount of reinforcement obtained after mutual defection was not enough to strengthen any response. When we tried using a higher level of *US* for mutual defection (which can reinforce the defection response), ZL defected frequently. Thus, the model predicted that if the amount of food that animals

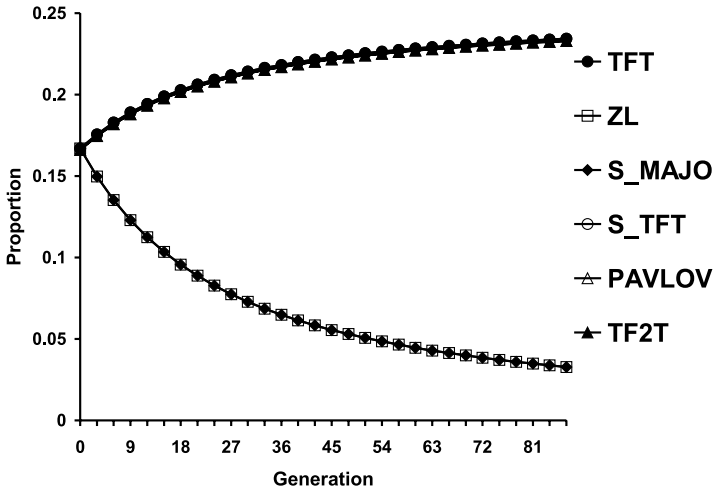

Figure 15. Group 2 schemes in the EPD for $pr = 0$ with ZL pretrained against all other strategies. In this case, ZL did not extinguish as in the experiment shown in Figure 7.

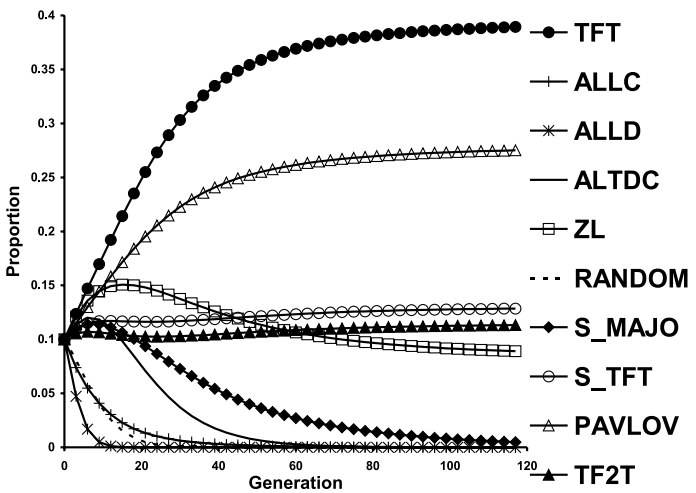

Figure 16. All schemes in the EPD for $pr = 0$ with ZL pretrained against all other strategies. The only strategies that survived were from group 2, except S_MAJO. In contrast with the results shown in Figure 11, ZL did not extinguish.

receive for mutual defection is enough for them, they will defect more frequently than predicted by the IPD. A more thorough study with real-time models of operant conditioning may provide new hypotheses for why the theoretical study of non-kin cooperation differs so much from that observed experimentally and give insight into the way experiments can be designed to show cooperation among animals.

7 Conclusion

Here it was shown that individuals having operant conditioning capacities can learn to play the prisoner's dilemma. In the iterated prisoner's dilemma, the operant condi-

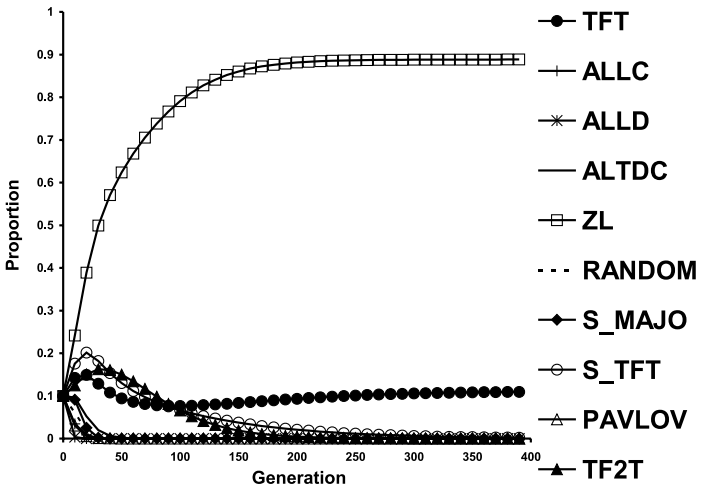

Figure 17. All schemes in the EPD for $pr = 0.01$ with ZL pretrained against all other strategies. ZL had the highest proportion in the final population. The other survivor was TFT.

tioning model (ZL) learned to cooperate with all the strategies of group 2 (cooperative strategies), except with S_MAJO in the case of nonrandom moves ($pr = 0$). The results on the evolutionary prisoner dilemma showed that ZL only survives against these strategies if the moves are noisy. This is because the other strategies can achieve perfect cooperation, while ZL is always below its ideal score because it loses points while learning the best reply to the opponent. However, when random moves are allowed in the game, ZL showed a relatively low sensitivity to noise, and the initial advantage of the other strategies diminished. In this way, ZL is able to survive under these conditions. It was also shown that ZL survives in all the proposed arrangements when it is pretrained against all the other strategies. These results support the idea that, although learning has the important property of adapting to a variety of strategies, the time that learning takes represents a disadvantage, in evolutionary terms, compared with nonlearning strategies.

Acknowledgments

This work has been partially supported by the fellowship "Beca Interna de Investigación EGD Dr. Ambrosio Tognoni" from the Rotary Club of Buenos Aires. We thank Dr. H. Dopazo, Eng. R. Zelmann, and Eng. S. Lew for their helpful comments on this manuscript.

References

1. Arita, T., & Suzuki, R. (2000). Interactions between learning and evolution: The outstanding strategy generated by the Baldwin effect. In *Proceedings of Artificial Life VII* (pp. 196–205).
2. Ainslie, G. W. (1974). Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior, 21*, 485–489.
3. Amsel, A. (1992). *Frustration theory*. Cambridge, UK: Cambridge University Press.
4. Ashby, W. R. (1952, 1960). *Design for a brain*. New York: Wiley.
5. Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390–1396.

6. Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
7. Baldwin, J. M. (1896). A new factor in evolution. *American Naturalist, 30*, 441–451, 536–553.
8. Bates, J. F., & Goldman-Rakic, P. S. (1993). Prefrontal connections of medial motor areas in the rhesus monkey. *Journal of Comparative Neurolobiology, 336*, 211–228.
9. Bolles, R. C. (1969). Avoidance and escape learning: Simultaneous acquisition of different responses. *Journal of Comparative Psychology and Physiology, 68*, 355–358.
10. Boyd, R. (1989). Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *Journal of Theoretical Biology, 136*, 47–56.
11. Brembs, B. (1996). Chaos, cheating and cooperation: Potential solutions to the prisoner's dilemma. *Oikos, 76*, 14–24.
12. Clements, K. C., & Stephens, D. W. (1995). Testing models of non-kin cooperation: Mutualism and the prisoner's dilemma. *Animal Behaviour, 50*, 527–549.
13. Connor, R. C. (1986). Pseudo-reciprocity: Investing in mutualism. *Animal Behavior, 34*, 1562–1566.
14. Crespi, L. P. (1942). Quantitative variation in incentive and performance in the white rat. *American Journal of Psychology, 5,* 467–517.
15. Darwin, C. (1859). *The origin of species*. London: John Murray.
16. Dragoi, V., & Staddon, J. E. R. (1999). The dynamics of operant conditioning. *Psychological Review, 106*, 20–61.
17. Dugatkin, L. A. (1997). *Cooperation among animals. An evolutionary approach*. Oxford Series in Ecology and Evolution. Oxford, UK: Oxford University Press.
18. Flood, M., Lendenmann, K., & Rapoport, A. (1983). 2 × 2 games played by rats: Different delays of reinforcment as payoffs. *Behavioral Science, 28*, 65–78.
19. Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neurophysiology of the frontal lobe* (p. 25). Philadelphia: Lippincott-Raven.
20. Gardner, R. M., Corbin, T. L., Beltramo, J. S., & Nickell, G. S. (1984). The prisoner's dilemma game and cooperation in the rat. *Psychological Reports, 55*, 687–696.
21. Green, L., Price, P. C., & Hamburger, M. E. (1995). Prisoner's dilemma and the pigeon: Control by immediate consequences. *Journal of the Experimental Analysis of Behavior, 64*, 1–17.
22. Grossman, K. E. (1973). Continuous, fixed-ratio, and fixed-interval reinforcement in honey bees. *Journal of the Experimental Analysis of Behavior, 20*, 105–109.
23. Goldman-Rakic, P. S. (1987). *Circuitry of primate prefrontal cortex and regulation of behavior by representational memory*. In F. Plum (Ed.), *Handbook of physiology: The nervous system* (pp. 373–417). Bethesda, MD: American Physiology Society.
24. Hamilton, W. D. (1964). The genetical evolution of social behavior I. *Journal of Theoretical Biology, 7*, 1–16.
25. Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.
26. Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of Experimental Animal Behavior, 4*, 267–272.
27. Herrnstein, R. J. (1969). Method and theory in the study of avoidance. *Psychological Review, 76*, 49–69.
28. Herrnstein, R. J., & Hineline, D. H. (1966). Negative reinforcement as shock-frequency reduction. *Journal of the Experimental Analysis of Behavior, 9*, 421–430.
29. Hull, C. L. (1943). *Principles of behavior*. New York: Appleton Century-Crofts.

30. Kacelnik, A., Kreps, J. R., & Ens, B. (1987). Foraging in a changing environment: An experiment with starlings. In M. L. Commons, A. Kacelnik, & S. L. Shettleworth (Eds.), *Quantitative analyses of behavior VI: Foraging* (pp. 63–87). Hillsdale, NJ: Erlbaum.

31. Kamin, L. J. (1957). The gradient of delay of secondary reward in avoidance learning. *Journal of Comparative and Physiological Psychology, 50*, 445–449.

32. Lew, S. E., Wedemeyer, C., & Zanutto, B. S. (2001). Role of unconditioned stimulus prediction in the operant learning: A neural network model. In *Proceeding of IEEE Conference on Neural Networks* (pp. 331–336).

33. Mackintosh, N. J. (1974). *The psychology of animal learning*. San Diego, CA: Academic Press.

34. Macy, M. W., & Flache, A. (2002). Learning dynamics in social dilemmas (2002). *Proceedings of the National Academy of Sciences of the United States of America, 3*, 7229–7236.

35. Mazur, J. E. (1984). Test of an equivalence rule for fixed and variable reinforcer delays. *Journal of Experimental Psychology: Animal Behavior Processes, 10*, 426–436.

36. Mazur, J. E. (1987). *An adjusting procedure for studying delayed reinforcement*. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior. Vol 5: The effect of delay and intervening events on reinforcement value* (pp. 55–73). Hillsdale, NJ: Erlbaum.

37. Mowrer, O. H. (1947). On the dual nature of learning, an interpretation of conditioning and problem solving. *Harvard Educational Review, 17*, 102–148.

38. Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature, 364*, 56–58.

39. Overmier, J. B., & Seligman, M. E. P. (1967). Effects of inescapable shock upon subsequent escape and avoidance responding. *Journal of Comparative and Physiological Psychology, 63*, 28–33.

40. Pear, J. J. (2001). *The science of learning*. Philadelphia: Psychology Press.

41. Pycock, C. J., Kerwin, R. W., & Carter, C. J. (1980). Effect of lesion of cortical dopamine terminals on subcortical dopamine receptors in rats. *Nature, 286*, 74–76.

42. Rachlin, H., & Green, L. (1972). Commitment, choice and self-control. *Journal of the Experimental Analysis of Behavior, 17*, 15–22.

43. Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts.

44. Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron, 35*, 395–405.

45. Sandholm, T., & Crites, R. H. (1995). Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems Journal, 37*, 147–166.

46. Schmajuk, N., & Zanutto, B.S. (1997). Escape, avoidance and imitation: A neural network approach. *Adaptive Behavior, 6*, 63–129.

47. Schmajuk, N., Urry, D., & Zanutto, B.S. (1998). The frightening complexity of avoidance: A neural network approach. In C. Wynne & J. Staddon (Eds.), *Models of action: Mechanisms for adaptive behavior*. Hillsdale, NJ: Erlbaum.

48. Schneirla, T. C. (1943). The nature of ant learning: II. The intermediate stage of segmental maze adjustment. *Journal of Comparative Psychology, 34*, 149–176.

49. Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36*, 241–263.

50. Schultz, W., Dayan P., & Montague, R. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593–1598.

51. Seligman, M. E. P., & Johnston, J. C. (1973). A cognitive theory of avoidance learning. In F. J. McGuigan & D. B. Lumsdenm (Eds.), *Contemporary approaches to conditioning and learning*. Washington, DC: Winston-Wiley.

52. Seligman, M. E. P., Rosellini, R. A., & Kozak, M. J. (1975). Learned helplessness in the rat: Time course, immunization and reversibility. *Journal of comparative and physiological psychology, 88*, 542–547.

53. Simpson, G. G. (1953). The Baldwin effect. *Evolution, 7*, 110–117.

54. Solomon, R. L., & Wynne, L. C. (1953). Traumatic avoidance learning: Acquisition in normal dogs. *Psychological Monographs, 67,* 354.

55. Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*(2), 135–170.

56. Staddon, J. E. R. (1983). *Adaptive behavior and learning*. Cambridge, UK: Cambridge University Press.

57. Staddon, J. E. R., & Zhang, Y. (1991). On the assignment-of-credit problem in operant learning. In M. L. Commons, S. Grossberg, & J. E. R. Staddon (Eds.), *Neural network models of conditioning and action*. Hillsdale, NJ: Erlbaum.

58. Staddon J. E. R., & Zanutto, B. S. (1997). Feeding dynamics: Why rats eat in meals and what this means for foraging and feeding regulation. In M. E. Bouton & M. S. Fanselow (Eds.), *Learning, motivation, and cognition:The functional behaviorism of Robert C. Bolles* (pp. 131–162). Washington, DC: American Psychological Association.

59. Staddon, J. E. R., & Zanutto, B. S. (1998). In praise of parsimony. In C. Wynne & J. Staddon (Eds.), *Models of action: Mechanisms for adaptive behavior*. Hillsdale, NJ: Erlbaum.

60. Stephens, D. W., & Clements, K. C. (1995). Game theory and learning: The law of effect and altruistic cooperation. In L. A. Dugatkin & H. K. Reeve (Eds.), *Advances in game theory and the study of animal behavior*. Oxford, UK: Oxford University Press.

61. Stephens, D. W., McLinn, C. M., & Stevens, J. R. (2002). Discounting and reciprocity in an interated prisoner's dilemma. *Science, 298*, 2216–2218.

62. Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplement, 2*, 8.

63. Thorndike, E. L. (1911). *Animal intelligence*. New York: Macmillan.

64. Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46,* 35–57.

65. Trivers, R. L. (1985). *Social evolution*. Menlo Park, CA: Benjamin Cummings.

66. Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*, 43–48.

67. Watanabe-Saguaguchi, K., Kubota, K., & Arikuni, T. (1991). Cytoarchitecture and intrafrontal connections of the frontal cortex of the brain of the Hamadryas baboon (*Papio hamadryas*). *Journal of Comparative Neurology, 311*, 108–133.

68. Watkins, C. J. (1989) *Learning with delayed rewards*. Ph.D. dissertation, Psychology Department, Cambridge University.

69. Wiener, N. (1948, 1961). *Cybernetics or control and communication in the animal and the machine*. Cambridge, MA: MIT Press.

70. Wilson, D.S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences of the U.S.A., 72*, 143–146.

71. Zanutto, B. S., & Lew, S. (2000). A neural network model of aversive behavior. In M.H. Hamza (Ed.), *Proceedings of the IASTED Neural Networks NN'2000* (pp. 118–123). Zürich: IASTED/ACTA Press.

Appendix A: Appetitive Learning

Thorndike [62, 63] was the first researcher to study systematically how nonreflex behavior can be modified as a result of experience. His experiments consisted of placing a hungry animal in a chamber. If the animal performed the appropriate response, the

door to the puzzle box would be opened, and the animal could exit and eat some food placed outside the door (appetitive reinforcement).

At the beginning of the experiments, the animals explored the chamber in a random way, until by chance the right response to exit was performed. To determine how a subject's behavior would change as a result of its experience, Thorndike would return an animal to the same puzzle box many times. He measured the amount of time it took the subject to escape on each trial. As trials progressed, the animal latency (the time between the presentation of the *CS* and the execution of the action) to escape gradually declined. He attributed this gradual improvement over trials to the progressive strengthening of a stimulus-response connection.

Appendix B: Aversive Learning

Solomon and Wynne [54] conducted an experiment showing many of the properties of aversive behavior. Their subjects were dogs, and the experiment consisted of a chamber with two rectangular compartments separated by a barrier several inches high where the subjects could move from one compartment to the other by jumping the barrier. Each compartment had a metal floor that could be electrified to deliver an electric shock, and two lights above the animal that could illuminate each compartment separately. In each trial, the light above the dog was turned off while the other light was turned on. If the dog remained in the dark compartment, after 10 seconds the animal received a shock until it jumped over the barrier. In this way, the animal could escape from the shock, but if it learned to jump before 10 seconds of the light being turned off, it could avoid the shock.

Solomon and Wynne measured latency as a function of the trial number. In the first trials, the latency was usually higher than 10 seconds; thus, the action performed was to escape from the shock. However, by the fifth trial the latency decreased below 10 seconds; thus, the action was to avoid the shock. They found that many dogs never again experienced a shock after their first avoidance response. This experiment posed a question that is called the avoidance paradox: How can the nonoccurrence of an event serve as a reinforcer for the avoidance response? This paradox led to the development of different theories of avoidance, such as the two-factor theory, the one-factor theory, and the cognitive theory.