Pergamon

Neural Networks, Vol. 10, No. 2, pp. 201–229, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(96)00067-6

CONTRIBUTED ARTICLE

A Dynamic Theory of Acquisition and Extinction in Operant Learning

VALENTIN DRAGOI

Duke University

(*Received* 2 *February* 1995; *revised and accepted* 3 *May* 1996)

Abstract—*This article offers a new neural network framework for understanding both the transients and the asymptotes of operant (instrumental) learning. The theory shows that interplay between simple short and long-term memory mechanisms is sufficient to explain a large number of operant phenomena. It describes short- and long-term effects of reinforcement and how these effects modulate the operant response, how novel events are detected and processed, and how their consequences also modulate the operant response. The critical features of the present theory are: (1) reinforcement expectancy is defined as the aggregate prediction of response-reinforcement and stimulus-reinforcement associations; (2) reinforcement expectancy controls the* rate of increase *of the operant response; (3) the response is controlled by a behavioral inhibition unit which integrates the mismatch between expected (long-term) and experienced (short-term) events. The model predicts the general features of several operant phenomena such as response selection, development of preference under different manipulations of reinforcement probabilities, negative contrast, partial reinforcement extinction effect, and spontaneous recovery. Implications of the present theory for other operant conditioning phenomena, classical conditioning, and avoidance behavior are suggested.*

Keywords—Assignment of credit, Contingency, Expectancy, Long-term memory, Operant conditioning, Recurrent choice, Reinforcement learning, Short-term memory.

1. INTRODUCTION

There are two kinds of experimental arrangement for the study of associative learning: classical and operant conditioning. The operations involved in classical conditioning comprise the repeated presentation of two classes of stimuli, a conditioned stimulus (CS) followed by an unconditioned stimulus (US), that results in the increase of a conditioned response (CR) when the CS alone is presented. In operant conditioning the reinforcement (equivalent to the US) is contingent. In both classical and operant learning the amount of conditioning is estimated by a change in the latency, speed, probability, or rate of occurrence of a specific response, when the delivery of a particular reinforcing event is made contingent upon the occurrence of the response. Skinner (1938) pointed out that there are three components in the operant conditioning contingency: (i) a discriminative stimulus which is present during a specific conditioning situation; (ii) the response; (iii) the stimulus following the response, i.e., the reinforcement. In the presence of the discriminative stimulus the reinforcement will occur if and only if the operant response occurs. Skinner called this relationship a *three-term contingency*, mainly referring to the issue of how stimuli that precede a behavior can control the occurrence of that behavior.

In a typical operant conditioning experiment each response on a specific cue, e.g., key pecking (pigeons) or lever pressing (rats), is reinforced on a probabilistic basis (variable-ratio reinforcement schedule, also known as VR schedule). The basic phenomena of VR responding (and associative learning in general) are acquisition and extinction. These issues are part of an

Acknowledgements: The author would like to thank Nestor Schmajuk, Armando Machado, Alliston Reid, Mark Cleaveland, as well as all the members of the Learning and Adaptive Behavior group for many helpful discussions and suggestions on these issues. He is especially grateful to his mentor, John Staddon, whose valuable comments contributed to a large extent to the present shape of the ideas advanced in this paper.

Research support was provided by grants to Duke University from National Science Foundation and the National Institute of Mental Health (J. E. R. Staddon, principal investigator).

Requests for reprints should be sent to: Valentin Dragoi, Department of Psychology: Experimental, Duke University, Box 90086 Durham, NC 27708, USA; Tel: (919)-660-5677; Fax: (919)-660-5726; e-mail: valentin@psych.duke.edu

important problem in operant conditioning, i.e., the *assignment of credit*[1]: when reinforcement is contingent on a particular response, the probability of that response generally increases; when reinforcement ceases, becomes less frequent or is presented independent of responding, response probability should decrease.

One interpretation of the assignment of credit problem is that it is the process by which animals "infer" causal relationships between reinforcement and response. To understand the basis of the inference process, learning theories are required to elucidate what is the active role of reinforcement on the dynamics of responding. As simple as it seems in principle this problem is actually hard to solve. One difficulty is that the same set of experimental manipulations can sometimes lead to different performance levels (response rate, response latency), depending on the actual moment of time when they are applied and on the past conditioning experience to which animals have been exposed. For instance, animals exposed to extinction show different degrees of resistance to extinction depending on the exact time when extinction is applied (Nevin, 1988). These historical properties of operant behavior suggest that a first step toward understanding the basis of conditioning would be to grasp the main aspects of the dynamics of a response, studied in isolation from other responses, with the reinforcement probability as the controlling variable (for instance in a simple VR schedule). This can be done by defining the *response unit* as a functional entity that sustains the behavior at the level of a *singular* operant response, i.e., the response on a specific cue.

The link between a singular response (CR in classical conditioning) and the input stimuli is made through a learning mechanism that should predict how the associations between the response or external stimuli and the reinforcement (response or stimulus associative strength) change in time. Most of the current conditioning models have in common the hypothesis, advanced first by Tolman (1932), according to which the organism generates expectancies or predictions of future events based on experienced reinforcement. In this respect, one important concept utilized by current theorists of Pavlovian conditioning is the notion of *expectancy* or *prediction*, a variable utilized to control directly the response. To further characterize the features of the present theory in the context of previous models of reinforcement learning it is necessary to introduce a brief historical survey on the recent evolution of concepts such as expectancy or prediction.

There have been earlier attempts to develop these concepts from the viewpoint of computational models of associative learning. For instance, in the Rescorla–Wagner model, (Rescorla & Wagner, 1972), learning occurs whenever the current US level differs from the current total reinforcement expectancy, when the total expectancy is defined as the sum of the associative strengths (a measure of CS–US associations) of all CSs present on the trial. Grossberg (1975, 1982) proposes that the processing of expected and unexpected event during conditioning is controlled by a novelty-detection mechanism that discriminates between the feedforward input signals and the feedback learned expectancy. Daly and Daly (1982) extended Rescorla and Wagner's concept of expectancy by incorporating Amsel's (1962) concept of frustration (frustration theory assumes that nonreinforcement in the presence of stimuli previously paired with reinforcement arouses an inferred aversive response which becomes associated to the stimuli present). They considered that the total reinforcement expectancy is the sum of three types of stimulus-specific associative strengths: approach strength (the US level is greater than the approach expectancy), avoidance strength (the US level is smaller than the avoidance expectancy), and counter-conditioning strength (animals learn to approach rather than avoid responses in aversive situations). Sutton and Barto (1981, 1990) implemented the first temporal-difference models of conditioning by building associations between *changes* in total reinforcement expectancy and eligibility traces (running average of recent values of each CS) that determine changes in the associative strengths of all stimuli present on the trial. Klopf (1988) defines the drive-reinforcement theory by proposing a learning mechanism that correlates *changes* in total reinforcement expectancy (changes in postsynaptic levels of activity) and *changes* in stimulus levels (changes in presynaptic levels of activity). As derived from the concept of reinforcement expectancy, Schmajuk (1995) develops the notion of novelty, defined by Pearce and Hall (1980) as the absolute difference between US expectancy and the actual US level. Novelty defined in this way drives an attentional system that controls the efficacy of the formation of S-S associations and modulates the strength of the CR.

In essence, I retain from these theories the idea that the response is controlled by the reinforcement expectancy, a variable which reflects the aggregate activity induced by the associative strengths of all stimuli present, including the response.

Following the definition of the response unit as an expectancy-driven functional entity, the second step in understanding operant behavior is to analyze the

[1] Samuel (1963) was the first to refer to this problem in the context of a checker-playing program where the task was to "assign credit" to moves that facilitate later moves that capture opponent pieces. Also, Minsky (1963) addressed the assignment of credit problem in the context of connectionist networks and animal learning.

interaction among response units that compete for reinforcement. The effects of this interaction can be studied under the framework of recurrent choice, i.e., the process by which animals "make decisions" or develop a certain preference in responding depending on the distribution of reinforcement between multiple choices (the word 'recurrent" suggests that the response is used as both output and input of a response unit). Recurrent choice is a popular topic in operant literature, especially since the mid-1950s, one reason being its generality, i.e., any action can be considered as the outcome of a choice process in that whenever the subject responds in reaction to a certain stimulus, it chooses in fact between non-responding and responding. Many animals (including humans) have been studied on a variety of operant procedures in which the allocation of behavior among choice alternatives (e.g., left or right) is measured as a function of the obtained rate of reinforcement from each alternative.

In choice, the theoretical emphasis so far has been based largely on molar equilibrium principles such as the *matching law* (Herrnstein, 1961), which states that under appropriate steady-state conditions the ratio of response rates, x/y, is approximately equal to the ratio of obtained reinforcement rates, $R(x)/R(y)$. However, despite their popularity, static matching relations say nothing about the specific mechanism underlying the process of choice on a moment-by-moment basis (real time) or at least on a response-by-response basis, instead they focus on the equilibrium states.

As opposed to molar models such as the matching law, molecular approaches attempt to predict both the steady state and the paths by which these equilibria are reached. In this respect, an important issue in operant learning is the study of the transients of acquisition and extinction with respect to different patterns of reinforcement. Surprisingly, excepting verbal theories available in operant conditioning, there are very few theories that offer (qualitative) explanations of operant phenomena in direct association with experimental data. Most of the existing operant conditioning models either concentrate on the mathematical apparatus at the expense of biological plausibility, thus accepting only loose connections with behavioral data, or they trade rigor for qualitative or quantitative fits to small data sets, e.g., Davis et al. (1993), Mazur (1992, 1995).

Most theories of choice use the notion of response "strength", as representing the output of a leaky integrator with fixed time-constant (Luce, 1995; Staddon & Zhang, 1991; Mazur, 1995). This is a way to encompass local effects of reinforcement by using a form of short-term memory for response and reinforcement events, however lacking the long-term event trace that embodies historical effects of reinforcement. The first attempt to build a real-time model that addresses directly the issue of long-term effects in choice is the theory proposed by Davis et al. (1993). Their assumption was that the animal calculates the overall probability of reinforcement for each response from the entire history of training, and the choice alternative with the highest probability of reinforcement is then selected on a winner-take-all basis. In essence, Davis et al. (1993) have implemented additive long-term memory units for responses and reinforcement by simply counting each response and each reinforcement.

Unfortunately, most of the existing theories of recurrent choice (and operant learning in general) test only a reduced number of behavioral paradigms. Therefore they cannot be used to explain extensive data sets, as would be required of a truly comprehensive model for instrumental learning. On the other hand, most of the expectancy-driven-response ideas mentioned earlier in the present section, utilized largely to develop comprehensive models of classical conditioning, have not been incorporated yet in models of operant learning (and it is not obvious the form in which these principles can be made useful to theoretical operant research). In this respect, the present article proposes a real-time neural network theory that relies heavily on the cooperation between short and long-term reinforcement expectancy mechanisms. The theory developed here offers plausible explanations for numerous adaptive effects of reinforcement learning as applied in operant studies, and proposes to identify and to describe the essential properties of acquistion and extinction within one unified framework.

In addition to the introduction, the article is divided into three more sections. Section 2 presents the basic hypotheses of the present theory and justifies from a behavioral point of view the underlying system of coupled differential equations. The section explains in a gradual fashion the general behavior of the network presenting the different components that interact dynamically. Section 3 presents the experimental data to be explained and shows computer simulations that mimic the experimentally observed behavior. Section 4 discusses other operant data sets that can be approached using the present framework, it suggests how selected classical conditioning and avoidance phenomena can be treated with the present formalism, and then compares the present theory with other existing reinforcement learning models. The final section summarizes the main findings of the paper.

2. THEORETICAL PRINCIPLES

This section introduces the theory of operant conditioning. I begin by describing the theory in functional terms, then I enumerate the main ideas.

The focus is on operant behavior in *transition*, i.e., the paths through which equilibria (stable states) are reached. Increasing evidence (e.g., Kacelnik et al., 1987; Mazur, 1992; Davis et al., 1993) suggests that in order to understand the specific process underlying acquisition and extinction in operant learning one should have a complete knowledge of the dynamic patterns of behavior in transition. This knowledge would help us to explain the topography of the steady state as well as allowing to say something about behavior on a moment-by-moment or on a response-by-response basis. In this way it would be possible to understand the influence of remote past history (the experimental conditions preceding the current one) on operant behavior.

Three issues about transients are addressed:

(a) The dependency of the *rate* of increase of responses during acquisition on the frequency of reinforcement. Since in operant conditioning the reinforcement depends on responding, it is necessary to estimate how the accumulation of contingent reinforcement affects the increase in responding. To solve this problem, I utilize the notion of *response–reinforcement associations*, first suggested by Mower (1960), according to which a specific response is selected by the reinforcement and consequently increases in strength. The response strength thus reflects the correlation between the response and the reinforcement. The accumulation of reinforced responses in the long run determines the formation of an expectancy of reinforcement at the level of that specific response. The reinforcement expectancy is utilized to control the rate at which the operant response increases, i.e., *the higher the reinforcement expectancy the faster the increase in response strength.*

(b) The animal's capability to react to *changes* in reinforcement contingency. To deal with this problem I suggest a simple system that detects variations in the strength of response–reinforcement associations by estimating *variations in reinforcement expectancy* as a comparison between expected and experienced reinforcement. Both expected reinforcement (long-term expectancy) and experienced reinforcement (short-term expectancy) are sensitive to the frequency of response–reinforcement associations and differ only by their rate of change in reaction to the occurrence of new events.

(c) The animal's capability to acquire *experience*. It is well known from the animal learning literature that an experienced animal is very different from a naive animal, even though the former lacked any substantial training for a period of many months. In this respect, it can be hypothesized that one structural change that characterizes animals exposed to reinforcement contingency (a change that should persists for a long time following the cessation of training) is the change in *learning rate*, i.e., the animal's capability to perform at a higher rate (speed of learning) in the presence of familiar cues despite a long absence from the experimental situation. Harlow (1949) called this improvement in the rate of learning by several different names, e.g., learning set or learning to learn. One way to achieve the improvement in the learning rate is by building slowly varying *associations between the long-term reinforcement expectancy and the response*. The previous reinforcement history ensures a certain level for the expectancy–response associations such that even when the reinforcement expectancy is zero (e.g., following prolonged extinction), once training is resumed the "experienced" animal shows an enhanced efficacy of increasing the *rate* of approaching the asymptote in responding (the response rate increases faster). In other words, what differs between an experienced and a naive animal is the strength of the connection between reinforcement expectancy and the response (the connection is stronger for an experienced animal by virtue of previous training), such that the input stimuli arriving at the same moment of time will be processed with different efficacy by a naive and an experienced animal.

In short, the present theory builds on the following ideas: (i) each reinforced response and discriminative stimulus increases its associative strength through facilitatory *response–reinforcement* (R–RF) associations and *stimulus–reinforcement* (S–RF) associations; (ii) competing responses and discriminative stimuli multiplied (gated) by their associative strength (R–RF and S–RF associations) constitute events stored as both short-term memory and long-term memory event-traces; (iii) the summed activity induced by response and stimulus events multiplied by their short or long-term memory event-traces determines the aggregate short-term reinforcement expectancy (a measure of expected reinforcement); (iv) the aggregate long-term reinforcement expectancy facilitates the operant response by modulating its *rate* of increase; (v) with extended training the efficacy of response control by the expected reinforcement increases via slow changes in the association between the reinforcement expectancy and the response (defined here as consolidation long-term memory); (vi) whenever the reinforcement is over-predicted (experienced reinforcement is smaller than expected reinforcement) a behavioral inhibition signal reduces the response strength.

2.1. The Model

Figure 1 shows a configuration with two mutually inhibitory response units (RU) that interact with the environment. The environment generates the reinforcement (RF), that is common for both response units, and the discriminative stimulus (S), that becomes

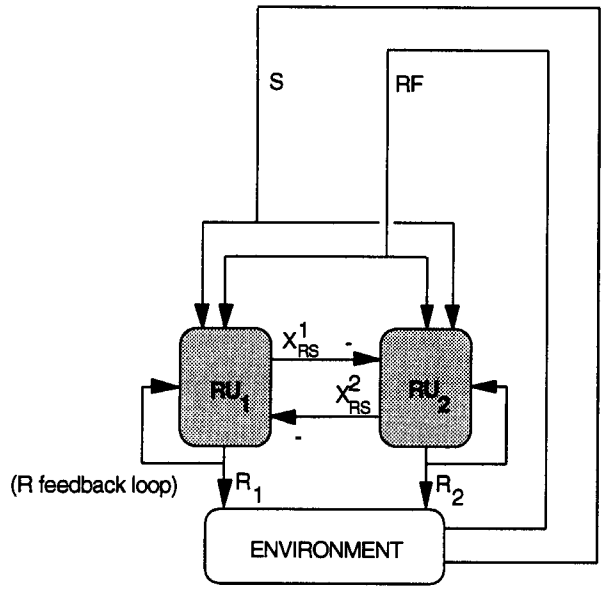

FIGURE 1. Interactions between response units (RU) and the environment. The environment generates the reinforcement (*RF*) which is common for both response units, and the discriminative stimulus (*S*) which becomes associated with the reinforcement at the level of both response units. The recurrent nature of the operant response (*R*) is represented by *R* feedback loops at the level of each response unit. The operant responses compete for reinforcing events through a process of mutual inhibition which is controlled by the response strength (X_{RS}).

associated with the reinforcement at the level of both response units. The recurrent nature of the operant response (R) is represented by R feedback loops at the level of each response unit. The operant responses compete for reinforcing events through a process of mutual inhibition which is controlled by the response strength (X_{RS}) of each RU, such that the response with the higher associative strength is selected by reinforcement.

Figure 2 shows the diagram of the model by presenting the detailed scheme of a response unit (RU). Each RU is characterized by a response-specific recurrent feedback loop with three inputs (the reinforcement, the discriminative stimulus, and the competing response) and one output (the actual response, R). In simple terms, the emitted response causes the formation of response-reinforcement (R–RF) associations by varying the connection strength w^R_{RF}. At the same time, the discriminative stimulus becomes associated with the reinforcement and changes its connection strength, w^S_{RF}. The response trace (X_{RT}) i.e., a running average of emitted responses, gated by the associative strength of the response, and the stimulus trace (X_{ST}) gated by the associative strength of the discriminative stimulus, drive the formation of LTM and STM memory traces for associations (w_{LM} and w_{SM}). The total

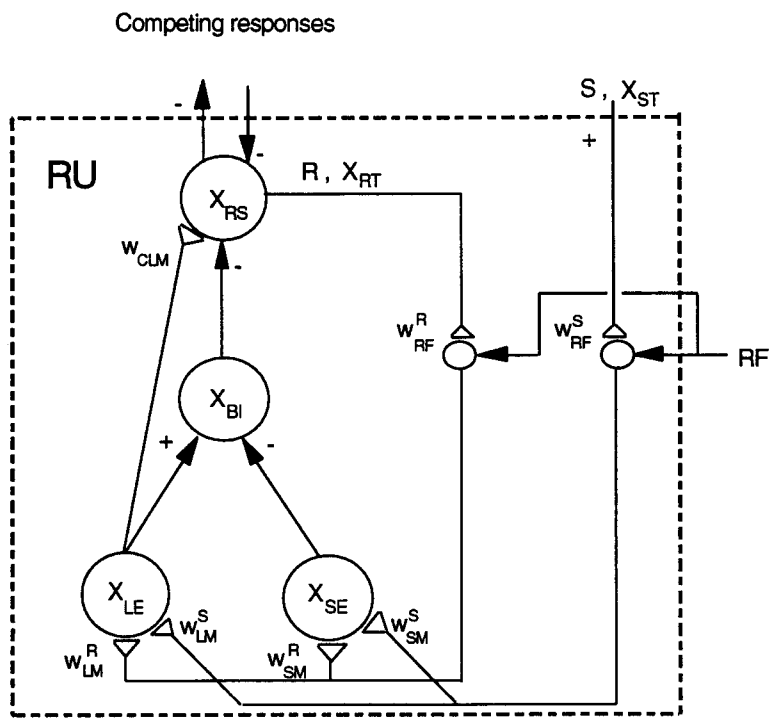

FIGURE 2. The diagram of a response unit. X_{RS}: *response strength;* *R*: response; X_{RT}: *response trace;* *S* discriminative stimulus; X_{ST}: *stimulus trace;* *RF*: reinforcement; w^R_{RF}: response associative strength; w^S_{RF}: stimulus associative strength; w^R_{SM}: short-term memory for *R–RF* associations; w^S_{SM}: short-term memory for *S–RF* associations; w^R_{LM}: long-term memory for *R–RF* associations w^S_{LM}: long-term memory for S–RF associations; X_{SE}: aggregate short-term reinforcement expectancy; X_{LE}: aggregate long-term reinforcement expectancy; w_{CLM}: consolidation LTM; X_{BI}: behavioral inhibition. (+) excitatory (fixed) connections, (−) inhibitory (fixed) connections; small triangles at nodes represent variable connections.

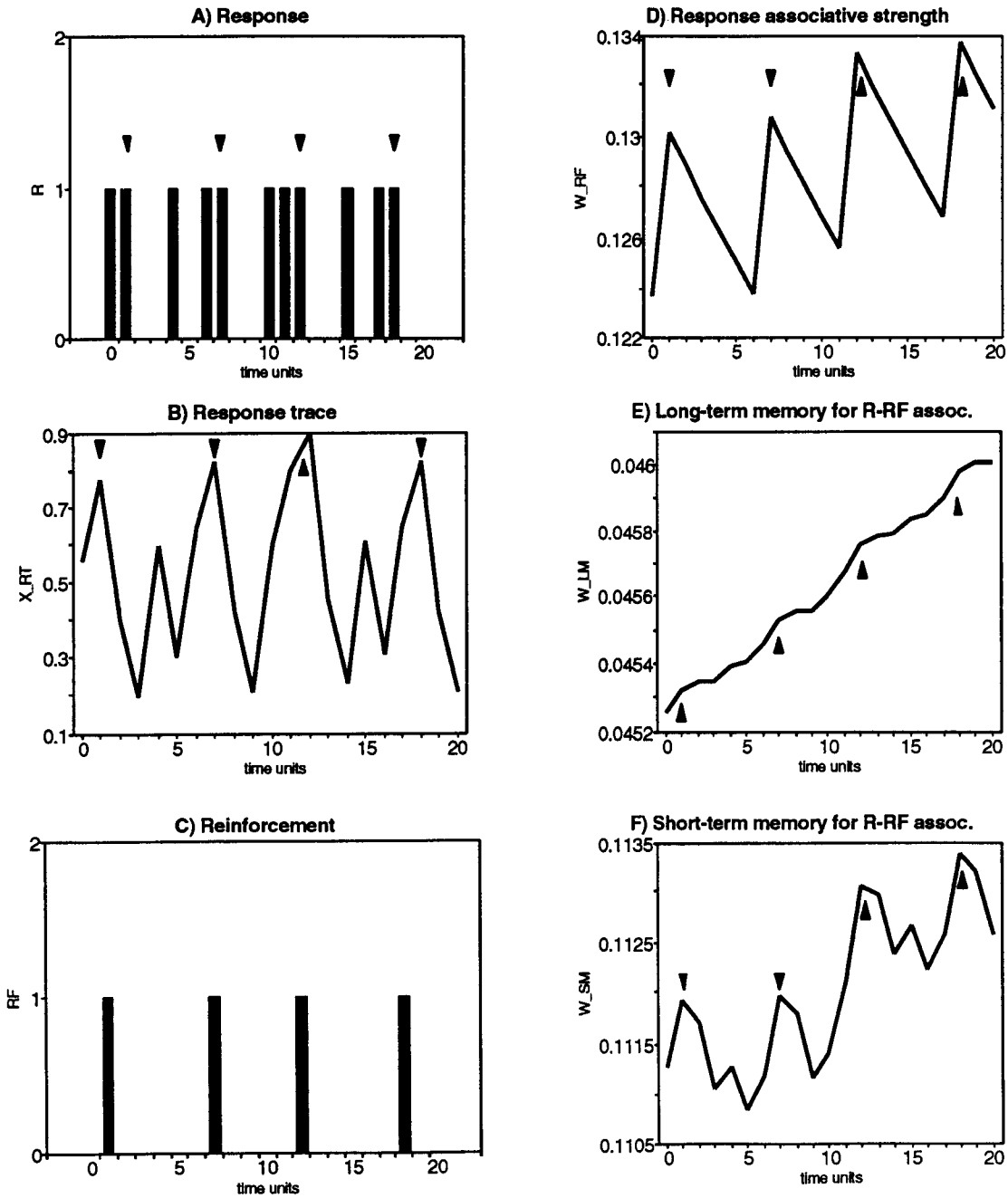

FIGURE 3. Illustration of model's dynamics during acquisition with FR 3 schedule of reinforcement. (A) The response sequence, *R*, generated during 20 time units (each reinforced response is marked with a black arrowhead). The graph does not show the response sequence before time 0. (B) The response trace, X_{RT}, dynamics. (C) The obtained reinforcement, *RF*, as a function of time. (D) The response associative strength, w_{RF}, dynamics. (E) The LTM trace for response-reinforcement associations, w_{LM}. (F) The STM trace for response–reinforcement associations, w_{SM}. (G) Upper half: the dynamics of the short and long-term *RF* expectancy, X_{LE} *and* X_{SE}; bottom half: the mismatch between long and short-term *RF* expectancy, X_{LE}–X_{SE}. (H) The behavioral inhibition, X_{BI}, increases whenever the aggregate long-term expectancy is greater than the aggregate short-term expectancy (for instance during nonreinforcement) and decreases otherwise. (I) The response strength, X_{RS}, increases with the accumulation of reinforcement. (J) Consolidation long-term memory, w_{CLM}, the slowly varying connection strength between the aggregate long-term expectancy and the response strength.

activity induced by the ongoing events which activate the LTM and STM traces constitute the aggregate long and short-term reinforcement expectancy (X_{LE} and X_{SE}). The response strength unit (X_{RS}) is facilitated by the aggregate long-term reinforcement expectancy unit and is inhibited by the behavioral inhibition (X_{BI}) signal, triggered by the mismatch between the aggregate long and short-term reinforcement expectancy units. The efficacy of response control by the aggregate long-term reinforcement

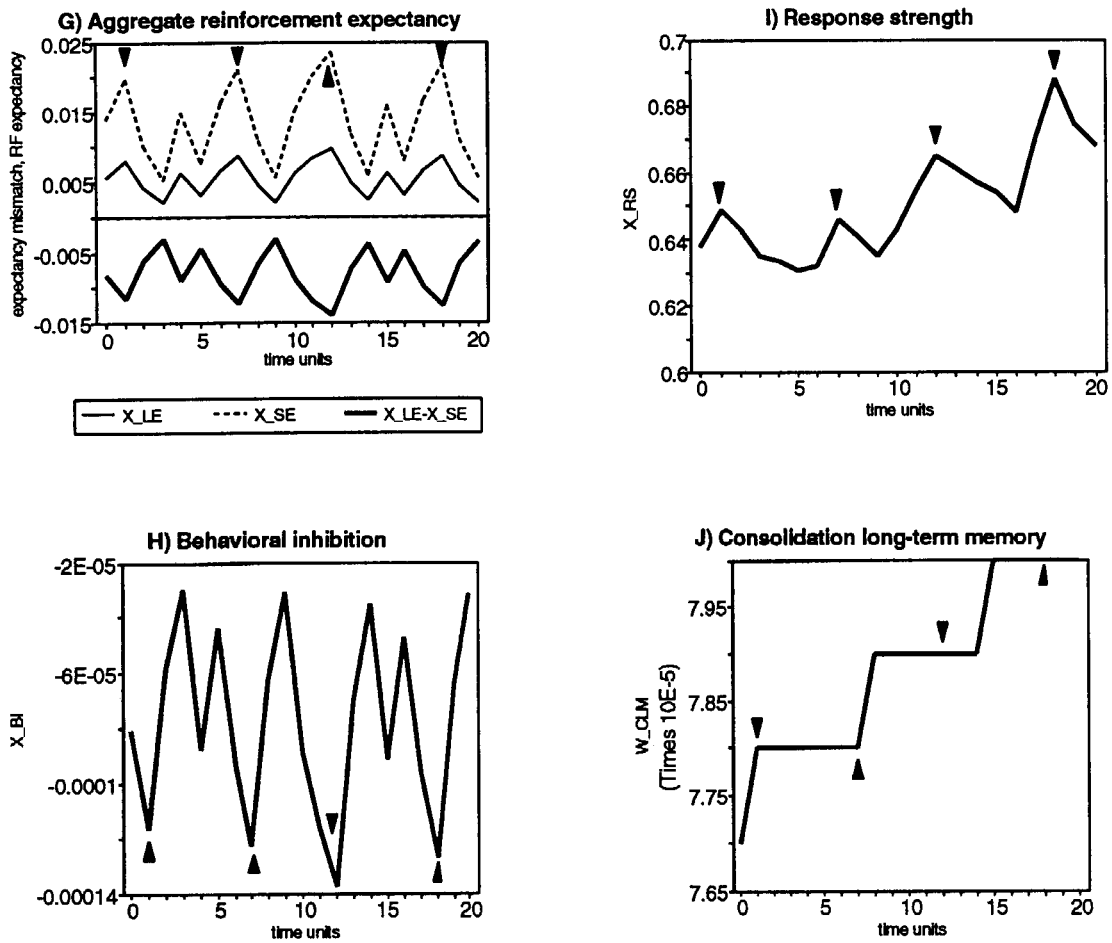

FIGURE 3 (continued)

expectancy is modulated by the consolidation LTM (w_{CLM}), i.e., the strength of the connection between X_{LE} and X_{RS}.

In order to understand the function of the model's various components, I present simulation results from (i) acquisition of partial reinforcement (Figure 3), and (ii) extinction (Figure 4). Figure 3A shows a response sequence generated during a fixed-ratio 3 schedule, or FR3 schedule (three responses are needed in order to receive one unit of reinforcement). The partially reinforced sequence is: 11001011001110010110 ("1" stands for an emitted response and "0" for no response), and spans between the time units 0 and 20 (the time unit when the reinforcement is applied is indicated by a black arrowhead). Each emitted response (R) activates the STM response trace, i.e., the (X_{RT}) representation. Figures 3B and 3C show the characteristic response trace dynamics (running average of emitted responses) and the obtained reinforcement (RF) as a function of time, i.e., the RF sequence 01000001000001000000100 ("1" stands for reinforcement awarded and "0" for no reinforcement). I hypothesize that the role of reinforcement is to enhance the associative strength of each response (w^{R}_{RF}) via a correlational (Hebbian) rule established between the response trace and the reinforcement. Figure 3D shows the time variation of the associative strength corresponding to the operant response. Notice that w^{R}_{RF} increases whenever the reinforcement is awarded in contiguity with the response (or relative contiguity in the case of delayed reinforcement) and decreases when the response is not reinforced. Because the response associative strength reflects the contiguity between the response trace and the reinforcement (and not between the short-lasting response and the reinforcement), it can be used as a basis for delayed conditioning. The magnitude of change in response associative strength is controlled by the frequency of *concurrent* response–reinforcement presentations, in other words by reinforcement contiguity.

Each response trace representation gated by its associative strength is viewed as an event which is stored as both long-term memory (LTM) and short-term memory (STM) representations, i.e., w_{LM} and w_{SM} respectively. Figure 3 (panels E and F) show the LTM and the STM traces for response–reinforcement

associations. In the present theory these temporal event traces differ in only one important respect: the rate of change of w_{SM} is high whereas the rate of change of w_{LM} is low. This behavior is clearly visible in Figure 3 (panels E and F): w_{LM} is perturbed to a lesser extent compared to w_{SM}, before and after each presentation of the reinforcement. Moreover, the high rate constant of the STM representations allows them to vary at a higher rate compared to the LTM representations, a fact also reflected in their magnitude. Notice also that the magnitude of STM and LTM representations (which reflect the response associative strength, w_{RF}^{R}) is influenced by response frequency, reinforcement frequency, reinforcement intensity, and reinforcement duration.

The summed activity induced by response and stimulus events gated by their short- and long-term memory event-traces determines the aggregate short-term reinforcement expectancy (X_{SE}) and the aggregate long-term reinforcement expectancy (X_{LE}), see Figure 2. Figures 3G shows the dynamics of X_{LE}, X_{SE}, and the dynamics of their difference, X_{LE}–X_{SE}. Notice that the dynamics of X_{LE} and X_{SE} are driven by the dynamic profile of response trace (this is because the response triggers the expectancy units). The magnitude of change in both X_{LE} and X_{SE} is proportional to the activity level of long- and short-term memory traces for R–RF associations (low for X_{LE} and high for X_{SE}).

Figure 3I shows that aggregate long-term reinforcement expectancy controls the rate of change of the response strength unit, X_{RS} (the response strength increases with the accumulation of reinforcement). The efficacy of this control is modulated by the slowly varying connection strength between X_{LE} and X_{RS} i.e., the consolidation long-term memory, w_{CLM} (see Figure 2). Figure 3J shows that w_{CLM} increases slowly following each reinforced response, by contrast to the dynamics of both w_{LM} and w_{SM}, much more sensitive to each reinforcement.

The occurrence of variations in reinforcement contingency is detected by the comparison between expected (long-term) events and experienced (short-term) events. If the organism over predicts the reinforcement (experienced events are worse than expected) then both X_{LE} and X_{SE} start decaying. Figure 3G illustrates the decay during the intervals following each reinforcement. Since the decay process happens at different rates (high for STM representations and low for LTM representations), after a certain number of non-reinforced responses the more persistent memory for expected reinforcement becomes more salient than the less persistent memory for experienced reinforcement. In this way, the change in the current reinforcement situation is detected by the behavioral inhibition (X_{BI}) unit which is driven by the difference between X_{LE} and X_{SE} (see Figure 2). Figure 3G, bottom half, illustrates the mismatch between the long and the short-term reinforcement expectancy. The output of the behavioral inhibition unit increases whenever both aggregate long-term and short-term reinforcement expectancies decrease (for instance during extinction), and decreases otherwise. Since the dynamics of behavioral inhibition are very important to understand the present theory, I will present them in two different situations: acquisition (Figure 3H) and extinction (Figure 4D).

It is easy to observe that X_{BI}'s profile shown in Figure 3H is out of phase with respect to the temporal dynamics of both X_{LE} and X_{SE}, and in phase with respect to the temporal dynamics of the difference $X_{LE} - X_{SE}$. As mentioned previously, the rate of change of the expectancy units is such that short-term expectancy varies faster than long-term expectancy, i.e.,

$$\left|\frac{dX_{LE}}{dt}\right| < \left|\frac{dX_{SE}}{dt}\right|.$$

When both expectancy units increase (the first time derivatives are positive) then

$$\frac{dX_{LE}}{dt} < \frac{dX_{SE}}{dt}$$

and therefore

$$\frac{d(X_{LE} - X_{SE})}{dt} < 0,$$

i.e., the expectancy mismatch decreases, a fact that contributes to the decrease in the amplitude of behavioral inhibition, see Figure 3 (panels G and H). When both expectancy units decrease (the first time derivatives are negative) then

$$-\frac{dX_{LE}}{dt} < -\frac{dX_{SE}}{dt}$$

and therefore

$$\frac{d(X_{LE} - X_{SE})}{dt} > 0,$$

i.e., the expectancy mismatch increases, a fact that favors the increase in behavioral inhibition (see Figure 3, panels G and H). The fluctuations in X_{BI} (which happen under partial reinforcement conditions, i.e., FR 3 schedule in our example) do influence the response strength dynamics in the sense that any increase in behavioral inhibition determines the waning of response strength (see Figure 3, panels H

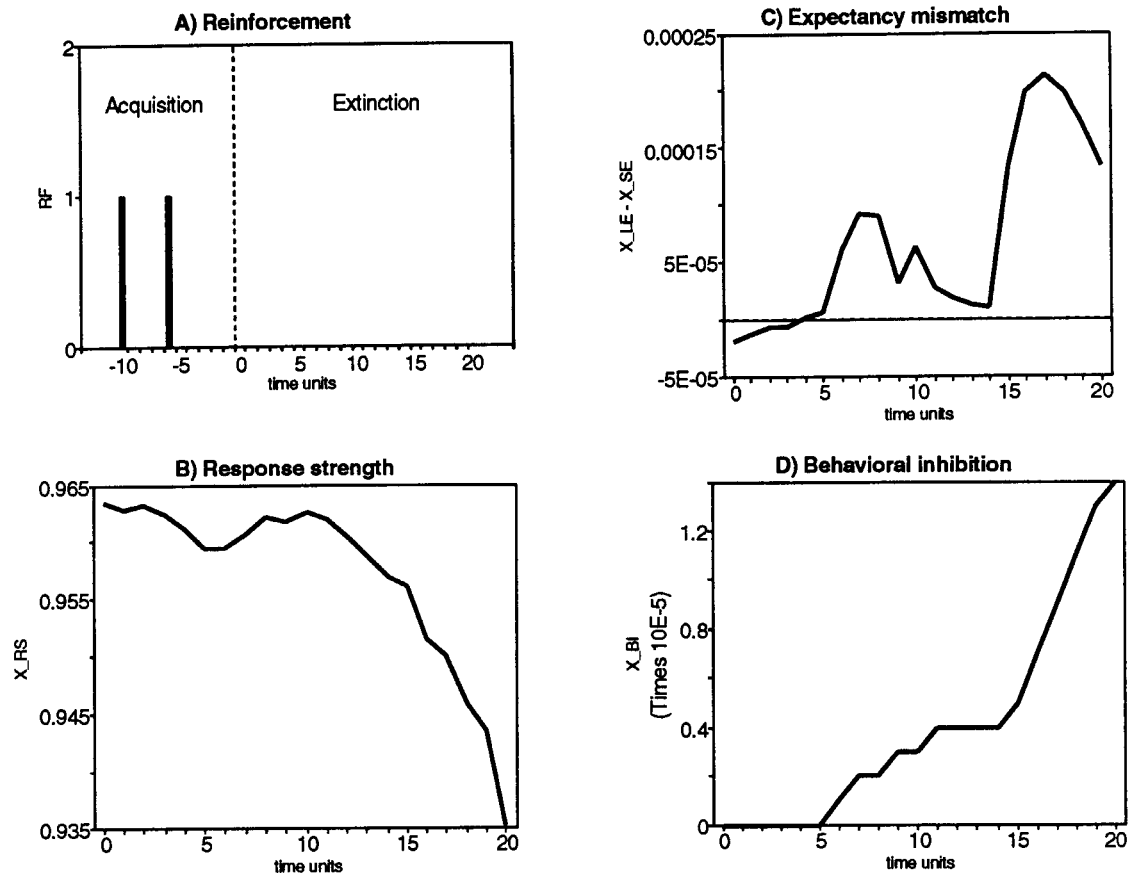

FIGURE 4. Illustration of model's dynamics during extinction. (A) The FR 3 schedule is in effect until time 0, the moment when extinction starts (*RF* is discontinued at time 0). (B) The response strength, X_{RS}, decreases due to the activity of the behavioral inhibition unit. (C) The change in expectancy mismatch, X_{LE}–X_{SE}. After sufficient time (after 5 time units) the decaying long-term expectancy becomes higher than the decaying short-term expectancy. (D) Behavioral inhibition, X_{BI}: as the expectancy mismatch becomes more positive (starting with time unit 5), X_{BI} increases and inhibits the response strength.

and I, where X_{BI} increases following each emission of RF, a fact that determines the decrease of X_{RS}).

Figure 4 shows the behavior of some of the model's compartments during extinction. Figure 4A shows that the last reinforcement is presented at time -6 (the FR 3 schedule is in effect until time 0, the moment when extinction starts). Figure 4B illustrates the decrease in response strength that matches the extinction situation (more detailed discussions on the effects of extinction will be presented in the next section). For clarity, Figure 4C omits to show X_{LE} and X_{SE} and illustrates only the change in the expectancy mismatch ($X_{LE} - X_{SE}$). Figure 4C shows that after sufficient time (after 5 time units) the decaying long-term expectancy becomes higher than the decaying short-term expectancy, even though they start decaying from different levels (low for X_{LE} and high for X_{SE}). Again, this result is a byproduct of the different sensitivities of the short and long-term memory units (w_{SM} and w_{LM}). Therefore, as the expectancy mismatch becomes more positive (starting with time unit 5), X_{BI} starts to increase and thus it inhibits the response strength, see Figure 4 (panels B and D).

In addition to the influence from the behavioral inhibition unit, the response strength can also be inhibited by competing responses (if they exist). I will analyze this situation (response competition) in one of the subsequent subsections (*Development of preference*) of this article.

2.2. Response Competition

Most of the existing theories of operant conditioning that deal with situations in which more than one response is involved assume the existence of separate representations for each response, e.g., Luce (1959), Davis et al. (1993), Schmajuk (1995). In accordance with these models of reinforcement learning, the present theory represents the alternate responses by response strength units, X_{RS_1} and X_{RS_2} in Figure 1. The level of the X_{RS_i} unit represents in real time the momentary preference for alternative i, where this alternative can be pecking key i, pressing lever i, etc.

To keep the analogy with the representation used by other theories, X_{RS} is functionally equivalent to the notion of V-value (Luce, 1959; Staddon & Zhang, 1991; Mazur, 1995; Davis et al., 1993).

In operant conditioning, every response followed by reinforcement (*RF*) increases in strength and becomes more likely to be re-emitted. Consequently, the probability of responding to the other alternatives (when they exist) decreases. In other words, if X_{RS_i} increases then X_{RS_j} should decrease, and vice-versa. A possible implementation of this response competition rule could be the process of lateral (mutual) inhibition. The present theory uses the mutual inhibition between the output (response) units, see Figure 1, to express response competition. The interresponse inhibition is mediated by fixed connections between the output units, X_{RS_i} and X_{RS_j} see eqn (1), where the term

$$-\alpha_2 X_{RS_i} \sum_{j \neq i} X_{RS_j}$$

represents the response inhibition.

$$\frac{dX_{RS_i}}{dt} + \alpha_1 X_{RS_i} = X_{LE_i}(w_1 + w_{CLM_i})(1 - X_{RS_i}) - \alpha_2 X_{RS_i} \sum_{j \neq i} X_{RS_j} - \alpha_3 X_{BI_i} X_{RS_i} \quad (1)$$

Here X_{LE_i} is the output of the aggregate long-term reinforcement expectancy unit, w_{CLM_i} is the consolidation LTM connection weight, X_{RS_j} is the output of the "j"th competing response strength unit, X_{BI_i} is the output of the behavioral inhibition unit; α_1 controls the spontaneous decay of X_{RS_i}, α_2 controls the strength of mutual inhibition between responses, and α_3 controls the strength of inhibition from X_{BI_i}; w_1 is the basal (fixed) connection between X_{LE_i} and X_{RS_i} (see Appendix).

The mutual inhibition process is behaviorally motivated by the observed positive and negative contrast effects (Crespi, 1952; Reynolds, 1961; Gutman 1977; Schwartz & Gamzu, 1977), that suggest that the extinction/elation of one alternative (or component of a multiple schedule) disinhibits/inhibits the response pattern to the other alternative (or component of a multiple schedule). The idea of competition between responses does not constitute a new assumption. For instance, Herrnstein (1970) views the rate of response in one component of a multiple schedule as depending on the rate of reinforcement in the adjacent components, supporting thus a process of comparison (competition) of the value of one component to its neighbors. Staddon and Hinson (1978) use behavioral competition as a mechanism for schedule interaction. Davis et al. (1993) use a winner-take-all rule to model response selection, also supporting the idea of response competition.

The next step in defining the operant response is to convert the content of the X_{RS_i} units into simple responses. In other words, one needs to establish how starting from a configuration of X_{RS_i} units a singular response can be generated. Being consistent with the assumption advanced by other models of reinforcement learning, e.g., Luce (1959), Mazur (1995), an individual response R_i is generated with the probability

$$p(R_i) = X_{RS_i} / \sum_j X_{RS_j}$$

where the sum from all X_{RS} units is taken. Here R_i is set to 1 if the subject responds to alternative i, according to the value of $p(R_i)$, and is 0 otherwise.

2.3. Short-term Memory (STM) Response Trace

The existence of STM response trace is justified by experimental results showing that animals can be conditioned to delayed reinforcement (Chung & Herrnstein, 1967; Killeen, 1968, 1970; McEwen, 1972). This suggests that the effect of each response remains active for an interval that varies with respect to response intensity and response duration. The present theory assumes that each response has a fixed intensity (equal to 1) and a fixed duration (one time unit). Making the analogy with the stimulus trace hypothesis (Hull, 1943; Grossberg, 1975; Sutton & Barto, 1981, 1990) I hypothesize that the trace of each response increases over time to a maximum following the emission of a response, and then decays to zero, see eqn (2).

$$\frac{dX_{RT_i}}{dt} = \alpha_4 (R_i - X_{RT_i}) \quad (2)$$

where R_i and X_{RT_i} represent the response and the response trace, and α_4 is the rate of increase and decay of X_{RT_i} (see Appendix). A similar equation can be written for stimulus trace, X_{ST_i}, where stimulus S replaces response R.

2.4. Response–Reinforcement (R–RF) and Stimulus-Reinforcement (S–RF) Associations

The present theory assumes that both responses and

[1] This type of equation that uses shunting excitation and inhibition is quite common in the neural networks literature (Grossberg, 1975).

stimuli enter the response units. At the response unit, reinforcement controls the strength of *R–RF* associations, w^R_{RF} and the strength of *S–RF* associations, w^S_{RF} (see Figure 2). Equation (3) expresses the changes in w^R_{RF}. It shows that *R–RF* associations are established when the response trace temporally overlaps with the reinforcement, such that w^R_{RF} reaches higher values when the interval between the onset of *R* and the onset of *RF* is smaller.

$$\frac{dw^R_{RF_i}}{dt} + \alpha_5 w^R_{RF_i} = \alpha_5 X_{RT_i} RF \tag{3}$$

Here X_{RT_i} is the trace of response R_i, *RF* is the reinforcement, and α_5 is the rate of increase and decay of w_{RF_i} (see Appendix). A similar equation can be written for stimulus associative strength, $w^S_{RF_i}$, where stimulus *S* replaces response *R*.

It is assumed that before the onset of conditioning response *R* has the weight w^R and stimulus *S* has the weight w^S. This is equivalent to the claim that depending on the animal's innate preference, each response and stimulus is characterized by a different initial capability to predict the reinforcement (w^R and w^S are loosely connected to the notion of preparedness from classical conditioning). At the end of multiple exposures to the reinforcement, response *R* will have the weight $w^R + w^R_{RF}$, and stimulus *S* will have the weight $w^S + w^S_{RF}$, quantities greater than w^R and w^S. During extinction, response *R* and stimulus *S* lose their predictive role, w^R_{RF} and w^S_{RF} decreasing to the level before conditioning, i.e., 0.

In a two-response situation, for instance recurrent choice, if R_1 occurs more frequently than R_2, the response choice R_1 becomes more strongly linked with *RF* compared with R_2, and therefore, through conditioning, $w^R_{RF_1}$ becomes greater than $w^R_{RF_2}$. The fact that different responses acquire different associative strengths (depending on the frequency of reinforcement contingent on particular responses) is used to solve the operant assignment of credit problem (Staddon & Zhang, 1991).

2.5. Memory Units for Event Processing

Consistent with the theoretical research in classical conditioning, where mechanisms of short and long-term memory proved to be successful in investigating the balance between the processing of expected and unexpected events (Grossberg, 1982; Sutton & Barto, 1990; Schmajuk, 1995) I propose to analyze whether simple short and long-term memory mechanisms can also be used in operant research. In this respect, the present theory assumes the storage of the response and stimulus trace gated by their association with the reinforcement, i.e., $X_{RT}(w^R + w^R_{RF})$ and X_{ST} ($w^S + w^S_{RF}$) as short and long-term memory traces ($w^R_{SM} - w^S_{SM}$ and $w^R_{LM} - w^S_{LM}$ in Figure 2). The difference between short and long-term lies in the rate of change of the memory units, i.e., w_{SM} varies faster than w_{LM}.

2.5.1. *Short-Term Memory Event Representations.* Short-term memory for *R–RF* and *S–RF* associations, w^R_{SM} and w^S_{SM} is a measure of the currently experienced correlation between *R* and *RF* and between *S* and *RF*. It consists of a set of connection weights with small time constant which increase every time that a new *R–RF* or *S–RF* association is formed, and then decrease until the occurrence of the following association [see eqn (4)]. If response *R* and stimulus *S* occur in the absence of *RF* (assuming that $w^R_{RF} = w^S_{RF} = 0$), they can still be processed by w^R_{SM} and w^S_{SM}, depending on the animal's response and stimulus-dependent predisposition to form new associations, labeled w^R and w^S in the model, see eqn (4). If response *R* and stimulus *S* are followed by *RF*, w^R_{RF} and w^S_{RF} increase and enhance the signals $X_{RT}(w^R + w^R_{RF})$ and $X_{ST}(w^S + w^S_{RF})$ that reach the memory units. Equation (4) shows that the changes in $w^R_{SM_i}$ are driven by the "presynaptic" potential that is proportional to $X_{RT_i}(w^R_i + w^R_{RF_i})$,

$$\frac{dw^R_{SM_i}}{dt} + \alpha_6 w^R_{SM_i} = \alpha_6 X_{RT_i}(w^R_i + w^R_{RF_i}) \tag{4}$$

where X_{RT_i}, is the trace of R_i, $w^R_{RF_i}$ is the associative strength of R_i, w^R_i is the basal (fixed) level of the connection between X_{RT_i} and X_{SE_i}, and α_6 controls the rate of increase and decay of $w^R_{SM_i}$ (see Appendix). A similar equation can be written for the STM for *S–RF* associations, $w^S_{SM_i}$, where stimulus *S* replaces response *R*.

2.5.2. *Long-term Memory Event Representations.* Long-term memory for *R–RF* and *S–RF* associations, w^R_{LM} and w^S_{LM}, is a measure of expected reinforcement. Similar to the short-term memory event representations, it consists of a set of connection weights which increase every time that a new *R–RF* or *S–RF* association is formed, and then decrease until the occurrence of the following association. Equation (5) shows that the changes in $w^R_{LM_i}$ are driven by the "presynaptic" potential that is proportional to $X_{RT_i}(w^R_i + w^R_{RF_i})$,

$$\frac{dw^R_{LM_i}}{dt} + \alpha_7 w^R_{LM_i} = \alpha_7 X_{RT_i}(w^R_i + w^R_{RF_i}) \tag{5}$$

where α_7 controls the rate of increase and decay of $w^R_{LM_i}$, with $\alpha_7 << \alpha_6$ (see Appendix). A similar equation can be written for the LTM for *S–RF*

associations, $w^S_{LM_i}$ where stimulus S replaces response R.

The similarity between w_{SM} and w_{LM} is that both reflect the strength of *R–RF* and *S–RF* associations. The difference between these two parallel memory units is the time course of their integration: short-term memory integrates events over a small time scale, whereas long-term memory integrates events over a more extended time scale.

2.6. Reinforcement Expectancy

Consistent with most theories of conditioning (e.g., Rescorla & Wagner, 1972; Grossberg, 1982; Schmajuk, 1995) I utilize the concept of reinforcement expectancy. Based on neuroanatomical and neurophysiological data, various neural theories of associative learning (e.g., Gray, 1982; Rudy & Sutherland, 1989; Schmajuk, 1995) proposed that reinforcement expectancy expresses the global effect of the associations between external stimuli (CSs) and the US. However, whereas these models define reinforcement expectancy as the aggregate prediction of the US, I define reinforcement expectancy as the aggregate prediction of response–reinforcement and stimulus–reinforcement associations.

At the level of a response unit, response and stimulus events multiplied by their associative strengths read out the corresponding short and long-term memory event-traces and determine the aggregate short-term reinforcement expectancy (a measure of experienced reinforcement) and the aggregate long-term reinforcement expectancy (a measure of expected reinforcement). In Figure 2 the output nodes, X_{SE} and X_{LE}, represent the aggregate short and long-term reinforcement expectancy that control the *RU*. Equation (6) expresses X_{SE_i}, as the algebraic sum of all the STM modules at the level of *RU*, i.e.,

$$X_{SE_i} = w^R_{SM_i} X_{RT_i}(w^R_i + w^R_{RF_i}) + \sum_{j=1}^{N} w^S_{SM_j} X_{ST_j}(w^S_j + w^S_{RF_j}) \tag{6}$$

where $w^R_{SM_i}$ is the STM trace of R_i, $w^S_{SM_j}$ is the STM trace of the discriminative stimulus S_j, $w^S_{RT_i}$ is the trace of R_i, X_{ST_j} is the trace of S_j, w^R_{RFi} is the associative strength of R_i, $w^S_{RF_j}$ is the associative strength of S_j, w^R_i is the basal (fixed) level of the connection between X_{RT_i} and X_{SE_i}, and w^S_j is the basal (fixed) level of the connection between X_{ST_j} and X_{SE_i}. Equation (7) expresses X_{LE_i} as the algebraic sum of all the LTM modules at the level of RU_i, i.e.,

$$X_{LE_i} = w^R_{LM_i} X_{RT_i}(w^R_i + w^R_{RF_i}) + \sum_{j=1}^{N} w^S_{LM_j} X_{ST_j}(w^S_j + w^S_{RF_j}) \tag{7}$$

where $w^R_{LM_i}$ is the LTM trace of R_i, $w^S_{LM_j}$ is the LTM trace of the discriminative stimulus S_j, X_{RT_i} is the trace of R_i, X_{ST_j} is the trace of S_j, $W^R_{RF_i}$ is the associative strength of R_i, $w^S_{RF_j}$, is the associative strength of S_j, w^R_i is the basal (fixed) level of the connection between X_{RT_i} and X_{LE_i} and w^S_j is the basal (fixed) level of the connection between X_{ST_j} and X_{LE_i}.

The X_{LE_i} unit sends direct excitation to the response strength unit, X_{RS_i}, such that it controls its rate of increase, i.e., the higher the level of X_{LE_i} the faster X_{RS_i} increases, sustaining thus a faster acquisition, see also equation (1).

2.7. Consolidation LTM

Equation (1) shows that the aggregate long-term reinforcement expectancy unit, X_{LE}, excites X_{RS} via a multiplicative term (gating LTM action), i.e., the consolidation LTM (w_{CLM} in Figure 2). The dynamics of w_{CLM} characterizes the history of reinforcement throughout training for each response unit. Equation (8) shows that the consolidation LTM is updated using a Hebbian rule applied to X_{LE_i} and X_{RS_i}

$$\frac{dw_{CLM_i}}{dt} + \alpha_8 w_{CLM_i} = \alpha_9 X_{LE_i} X_{RS_i} \tag{8}$$

where α_9 and α_8 are the rates of increase and decay of w_{CLM} (see Appendix). The rate at which w_{CLM} decays during extinction is lower than the rate of acquisition. Notice that w_{CLM} changes slowly in time, at a much slower rate than w_{LM}, i.e., $\alpha_8, \alpha_9 << \alpha_7 << \alpha_6$. As a consequence of the time course of w_{CLM}'s integration, the effects of its variation become important only after long training sessions.

2.8. Behavioral Inhibition

Gray (1971), following the tradition of Sokolov (1960), suggested that a behavioral inhibition system, activated by signals of punishment or nonreward, innate fear stimuli, or novel stimuli, generates a behavioral inhibition signal which reduces the ongoing behavior. Extending these ideas, Gray (1982) developed a qualitative neural theory of information processing in the septo-hippocampal system according to which the subiculum acts as a comparator (Vinogradova, 1975) between the US prediction and the current events. If a mismatch is detected, a behavioral inhibition signal accesses the cingulate cortex (Swanson, 1978) allowing the septo-hippocampal system to inhibit the activity in the prefrontal cortex and the cerebellum. Schmajuk (1995) builds on this framework and uses the mismatch between the actual and predicted intensity of the US to control behavioral

inhibition (during avoidance. In line with all of these approaches, I suggest the use of a behavioral inhibition unit (X_{BI} in Figure 2) whose main role is to detect variations in reinforcement contingency as a result of the different time constants for the aggregate short and long-term reinforcement expectancies. If the reinforcement conditions become better (larger size, smaller delay, increased duration, higher probability), the aggregate short-term reinforcement expectancy increases faster than the aggregate long-term reinforcement expectancy. If the reinforcement conditions become worse (smaller size, larger delay, decreased duration, lower probability), the aggregate short-term reinforcement expectancy decreases faster than the aggregate long-term reinforcement expectancy. The X_{BI} unit integrates the difference (mismatch) between expected (long-term) and experienced (short-term) reinforcement, i.e., $X_{LE} - X_{SE}$. Here, X_{BI} increases whenever *R–RF* and *S–RF* associations are overpredicted, i.e., $X_{LE} - X_{SE} > 0$, and decreases whenever *R–RF* and *S–RF* associations are underpredicted, i.e, $X_{LE} - X_{SE} < 0$. After both X_{LE} and X_{SE} become 0, X_{BI} slowly relaxes to 0. In essence, behavioral inhibition quantifies an "emotional" state (frustration/elation) that plays a crucial role during the occurrence of variations in reinforcement schedules. The present theory is similar in this sense to many other theories of reinforcement learning, such as Amsel's (1962) frustration theory, Daly and Daly's (1982) model, Grossberg's (1982) forward input–feedback expectation mismatch concept, or Schmajuk's (1995) behavioral inhibition.

Equation (9) shows that the difference between X_{LE_i} and X_{SE_i} drives the dynamics of behavioral inhibition,

$$\frac{dX_{BI_i}}{dt} + \alpha_1 X_{BI_i} = \alpha_{10}(X_{LE_i} - X_{SE_i})(1 - X_{BI_i}) \qquad (9)$$

where X_{LE_i} is the aggregate long-term reinforcement expectancy at the level of RU_i, X_{SE_i} is the aggregate short-term reinforcement expectancy at the level of RU_i, and α_1 and α_{10} are the rate constants of decay and increase of X_{BI_i} (see Appendix).

3. THE DYNAMICS OF OPERANT CONDITIONING

The aim of this article is to account for some of the major phenomena of operant learning (with emphasis on recurrent choice) from the viewpoint of the theory introduced in the above section. The experimental situations that are referred in this paper are based on probabilistic reinforcement schedules with animal subjects (mainly rats and pigeons) and food reinforcement. The data set is the real-time pattern of responses under different manipulations of reinforcement probability as a function of time. The analysis is restricted to simple and concurrent variable ratio (or probabilistic) schedules, where the pattern of preference in responding develops in time to an asymptote. In concurrent ratio schedules the asymptote is typically the same, although the rate at which the asymptote is reached during the transient phase is a function of the reinforcement probability. The properties of data that I attempt to explain are expressed as qualitative patterns of change in real-time responding. The different response patterns are correlated to the changes thus obtained.

In all the simulations presented in this section there are analyzed situations involving one or two operant responses, a reinforcement, and a discriminative stimulus. It is also possible to deal with multiple responses, i.e., each response has associated a specific *RU* (see Figure 1), and the RUs mutually inhibit each other. For instance, I carried out computer simulations (not described here) which analyze situations involving five and ten operant responses that were reinforced with different probabilities, showing that the response reinforced at the highest rate is selected by reinforcement (the values of all parameters are kept constant).

Whenever the computer simulations start, the activity level of all the variables is initialized to 0, with the exception of the response strength units which are initialized to 0.5. The reinforcement conditions are signaled by the discriminative stimulus which switches from 0 to 1 and triggers one response at random. The next response, R_i, is generated with the probability

$$p(R_i) = \frac{W_{RS_i}}{\sum_{j=1}^{N} X_{RS_j}}$$

from the set of all N responses that are available. If the response R_i is generated, a random number between 0 and 1 is compared with the reinforcement probability for R_i. If the random number is smaller or equal to the reinforcement probability, a reinforcement is set to 1 for one time unit (the same method used in experimental conditions during a concurrent VR-VR schedule).

Some of the most important phenomena of operant conditioning are addressed below.

3.1. Response Selection

Experimental data

Response selection can be analyzed in a simple two-armed bandit situation. In this situation (concurrent VR-VR schedule) responses on a rich and a lean side are paid off with different probabilities (higher for the

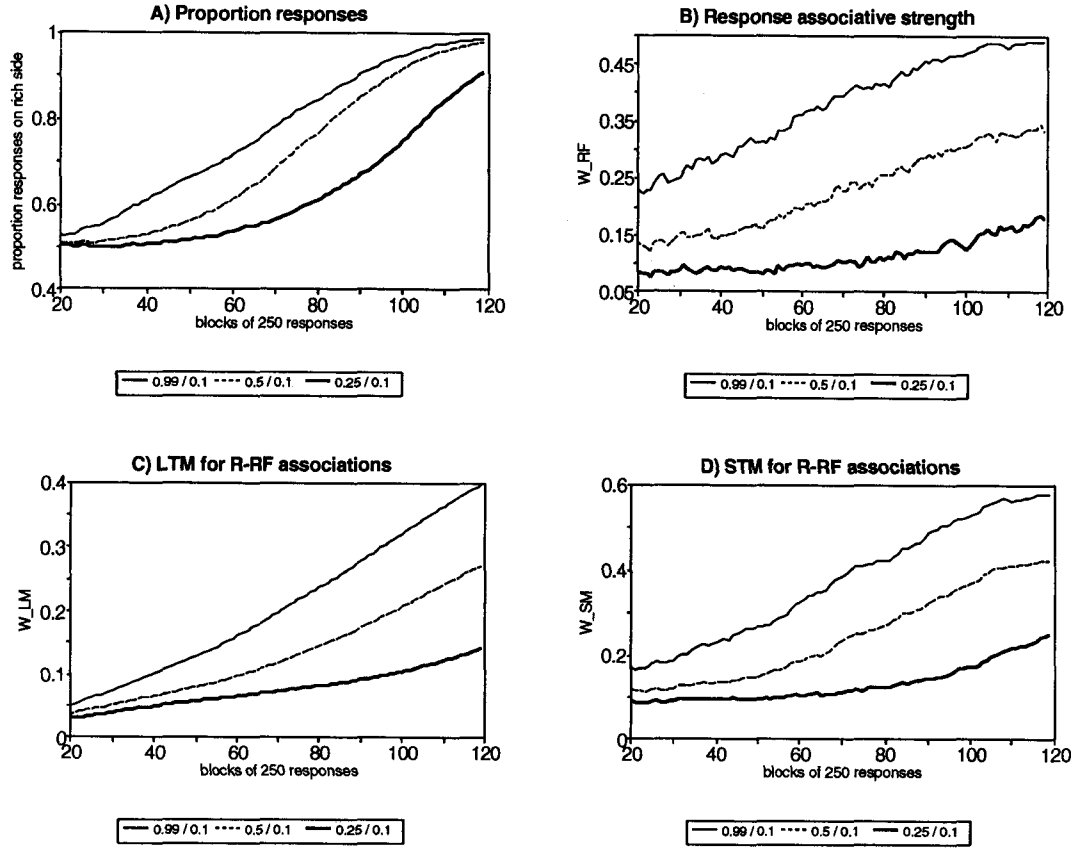

FIGURE 5. Response selection. (A) Proportion of responses on the rich side as a function of blocks of 250 responses. Responses reinforced at a higher rate are selected faster by the reinforcement (choice becomes more exclusive at higher *RF* rates). (B) Responses reinforced at a higher rate tend to develop stronger associative strengths, w_{RF}. (C–D) The rate of increase of STM and LTM for response-reinforcement associations, w_{LM} and w_{SM}, is positively correlated with the reinforcement probability.

rich side) and preference develops toward the rich side (effect known as exclusive preference), a situation in which all response ratios approach unity at a rate controlled by the reinforcement probability for the rich side. The problem can be stated as follows (Staddon, 1983): (i) how does the organism select the "rich" response without an explicit *a priori* "knowledge" of what that response is? (ii) why does the speed of response selection act at a higher rate when the reinforcement probability on the side increases?

Simulation results

To answer these questions, I consider a generic situation in which three different conditions are run under a concurrent VR-VR schedule. The lean side provides reinforcement with fixed probability, 0.1, while the rich side is set at three different levels, i.e., 0.99, 0.5, and 0.25 Figure 5A shows the dynamics of the proportion of responses on the rich side in each of these conditions as a function of blocks of 250 responses. As the reinforcement probability of the rich side increases, the rate of response selection also increases. To explain this result, it is assumed that initially the subject samples both alternatives with equal probability, i.e., both X_{RS} units are set to the same level. Figure 5B shows the profile of the response associative strength on the rich side, from which it can be concluded that responses reinforced at a higher rate tend to form stronger associations (w^{R}_{RF}) with the reinforcement. Figure 5 (panels C-D) shows that the STM and LTM traces for *R–RF* associations at the level of the rich side also increase, causing the aggregate expectancy units (X_{SE} and X_{LE}) to grow (the rate of increase of STM and LTM for *R–RF* associations is proportional to the reinforcement probability). Since X_{LE} excites the response strength unit by controlling its rate of increase, the "rich" response is selected faster with the increase in the aggregate long-term reinforcement expectancy. At the same time, the competition between operant responses ensures that the strength of responses on the lean side wanes, whereas responses on the rich side gradually become more vigorous. In this way, after a sufficient number of responses the subject fixates on the rich side (Figure 5A). In addition to the reinforcement probability, the rate of fixation is also positively correlated to the strength of inhibition between the two response units, but this result is not

analyzed (the present simulations use a fixed strength of inter response inhibition).

3.2. Development of Preference

In order to understand how preference develops in operant learning, one should probably begin with the analysis of simple choice situations such as concurrent variable ratio schedules, VR-VR. It is clear that under these conditions the preference develops toward the side with the higher reinforcement probability. The principal question to be addressed is how the distribution of the two reinforcement probabilities affects the rate of transition toward the "winning" alternative. In this respect, I distinguish between two different situations that suggest the following determinants of the development of preference: (i) absolute difference between reinforcement probabilities—the two probabilities vary such that their ratio is held constant; (ii) ratio between reinforcement probabilities—the two probabilities vary such that their absolute difference is held constant.

Experimental data. Bailey and Mazur (1990) suggest that the development of preference for one alternative depends on the discriminability of the two alternatives, showing that with concurrent VR-VR schedules the acquisition of preference occurs more rapidly with larger ratios between the probabilities of reinforcement, even when their absolute difference is held constant. For instance, suppose that in one condition the two reinforcement probabilities are 0.40 and 0.30, and in another condition the two probabilities are 0.12 and 0.02. Even though the difference between the two probabilities is 0.10 in both conditions, the transition to preference for the higher probability of reinforcement is much faster in the 0.12/0.02 condition (ratio 6) than in the 0.40/0.30 condition (ratio 1.333).

In the same vein, Mazur (1992) showed that when the ratio between the two probabilities of reinforcement is held constant, the preference develops according to the absolute difference between the two probabilities. For instance, suppose that in one condition the two reinforcement probabilities are 0.16 and 0.08, and in another condition the two probabilities are 0.08 and 0.04. Even though the ratio between the two probabilities is 2 in both conditions, the transition to preference for the higher probability of reinforcement is much faster in the 0.16/0.08 condition (difference 0.08) than in the 0.08/0.04 condition (difference 0.04). These findings seem to point toward a Weber law effect in the acquisition of preference, i.e., two stimuli (reinforcement probabilities) are more easily discriminated (or processed) if they differ by a larger percentage or absolute difference.

Mazur's experimental results are inconsistent with most of the known models of acquisition, which are unable to predict which conditions would have rapid transition rates and which would have slow ones (Mazur, 1995). However, Grossberg (1972), in his analysis of punishment and avoidance effects, did derive a Weber law in the development of preference as an emergent property of a gated dipole opponent process coupled to conditioned reinforcers.

3.2.1. *Effect of Ratio between Reinforcement Probabilities.*

Simulation results. I have simulated the experiment described in Bailey and Mazur (1990). Each response on one key (rich) was reinforced with a probability p_1, and each response to the other key (lean) was reinforced with a probability $p_2 < p_1$. By keeping $p_1 - p_2$ fixed (equal to 0.10), while varying p_1/p_2, the development of preference is compared across three groups. The pairs of reinforcement probabilities are: 0.4/0.3, 0.2/0.1, and 0.12/0.02. Figure 6A shows simulation results that agree with experimental data, i.e., the preference develops faster with the larger ratios between the two reinforcement probabilities. Figure 6B shows that, for the situation described in the current simulation, even though lower reinforcement probability ratios are equivalent to higher reinforcement expectancies, preference is an increasing function of p_1/p_2.

This effect can be explained as a conjoint result of response competition and long-term reinforcement expectancy. According to the theory, the rate of increase of X_{RS_1} ("1" is the richer side) is influenced by two factors, see equation (1). One is excitatory, the term $X_{LE_1}(w_1 + w_{CLM_1})(1 - X_{RS_1})$ which sustains acquisition, and one is inhibitory, the term $-\alpha_2 X_{RS_1} X_{RS_2}$ which determines the strength of competition between responses. Consolidation LTM is not involved because it varies very slowly and the time course of Bailey and Mazur's experiment (one session) does not allow w_{CLM} to increase to a value which can influence acquisition. It has been previously shown that the long-term reinforcement expectancy (X_{LE}) varies proportionally to the LTM trace for *R–RF* associations (w_{LM}), and that w_{LM} varies proportionally to the frequency of contingent reinforcement. Therefore, the higher the ratio between the two reinforcement probabilities (p_1/p_2) the higher the ratio between the corresponding X_{LE} units (X_{LE_1}/X_{LE_2}) which sustain the two competing responses. Since the long-term reinforcement expectancy controls the rate of increase of the response, higher ratios between the X_{LE} units (corresponding to the rich and the lean sides) will determine higher ratios between the level of the corresponding

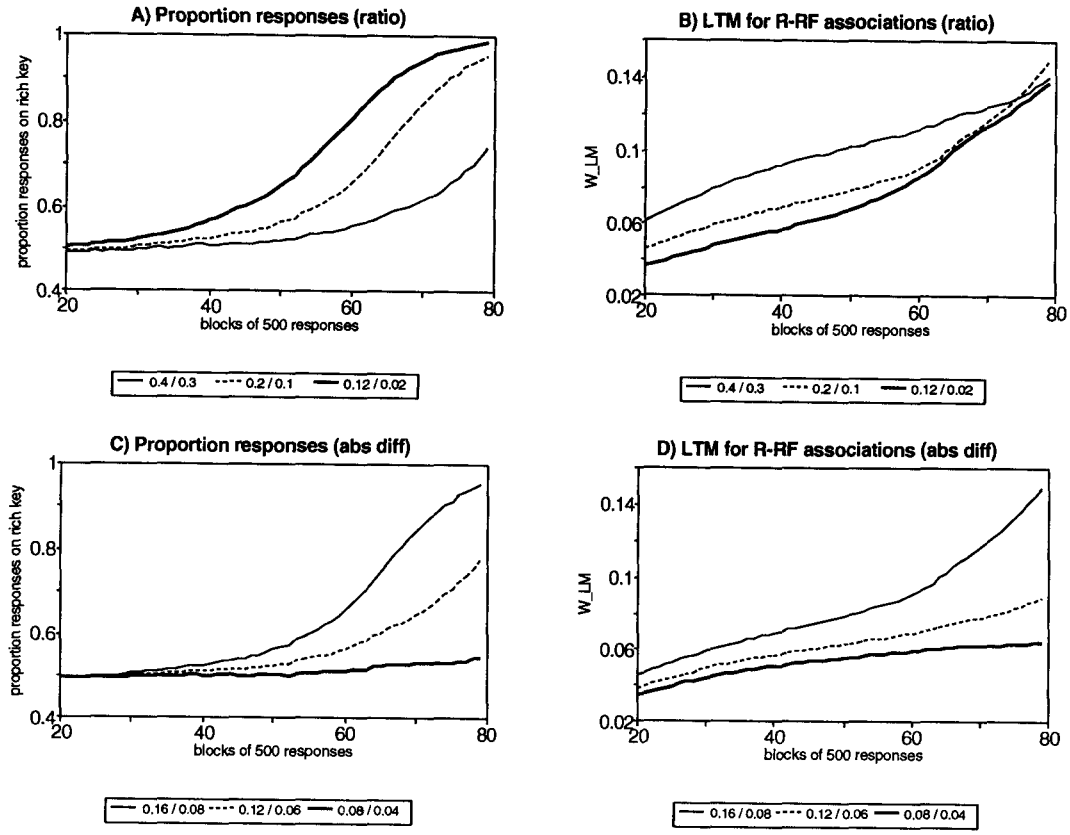

FIGURE 6. The influence of ratio and absolute difference between reinforcement probability on the development of preference. (A) Proportion responses on rich key as a function of blocks of 500 responses: preference develops faster with the larger ratio between the two reinforcement probabilities (the ratio 6 ensures the fastest fixation). (B) Long-term memory for *R–RF* associations, w_{LM}, for the rich key: even though lower probability ratios are equivalent to higher reinforcement expectancies, however preference is an increasing function of probability ratio. (C) Proportion responses on rich key as a function of blocks of 500 responses: preference develops faster with the absolute difference between the two reinforcement probabilities (the absolute difference 0.08 ensures the fastest fixation). (D) Long-term memory for *R–RF* associations, w_{LM}, for the rich key: lower absolute differences between reinforcement probability are equivalent to lower reinforcement expectancies.

response strength units (X_{RS_1}/X_{RS_2}). Notice that this dependency holds only in transition, not in the steady state of the X_{RS} units, as equation (1), which describes X_{RS}, is a saturating equation with respect to the term which contains X_{LE}. According to the response rule of the theory, a singular response, R_1, is emitted with the probability $p(R_1) = X_{RS_1}/(X_{RS_1} + X_{RS_2})$, a function that varies monotonically with X_{RS_1}/X_{RS_2}. In this respect, if the ratio p_1/p_2 increases, then X_{RS_1}/X_{RS_2} also increases determining a higher probability, $p(R_1)$, with which the response R_1 is generated, i.e., more frequent response on the rich side and less frequent responses on the lean side. Therefore, there will be less inhibition on X_{RS_1} by X_{RS_2}, resulting thus in a faster preference for the richer side. Notice that the preference for alternative 1 is defined as $X_{RS_1}/(X_{RS_1} + X_{RS_2})$.

3.2.2. *Effect of Absolute Difference between Reinforcement Probabilities.*

Simulation results. I have simulated the experiment described in Mazur (1992). Each response on one key (rich) was reinforced with a probability p_1, and each response to the other key (lean) was reinforced with a probability $p_1 < p_2$. By keeping p_1/p_2 fixed (equal to 2), while varying $p_1 - p_2$, the development of preference is compared across three groups. The pairs of reinforcement probabilities are: 0.16/0.08, 0.12/0.06, and 0.08/0.04. Figure 6C shows simulation results that agree with experimental data, i.e., the preference develops faster with the absolute difference between the two reinforcement probabilities. Figure 6D shows that lower absolute differences between reinforcement probabilities are equivalent to lower reinforcement expectancies.

Similar to the case of probability ratios, this effect can be explained as a conjoint result of response competition and long-term reinforcement expectancy. Given that the long-term reinforcement expectancy varies proportionally to the reinforcement probability, at equal ratios between reinforcement probabilities one should expect equal ratios between reinforcement expectancies. However, if the absolute difference

between reinforcement probabilities increases (at the same ratio), the probability of the rich side also increases. This leads to the formation of stronger *R–RF* associations at the level of the richer alternative, causing X_{LE} to increase in time to a higher level. This contributes to the facilitation of X_{RS_1} which increases faster and sends more inhibition to X_{RS_2}, sustaining thus a higher rate of fixation. In other words, Mazur's (1992) results can be explained by means of the excitatory effect of the absolute reinforcement probability for the rich alternative.

3.3. Contrast Effects

Experimental data. The term contrast effect refers to those situations in which exposure to more than one reinforcement condition exaggerates the difference between the performance maintained under each condition alone. If subjects are switched from a CRF (continuous reinforcement, i.e., every response is reinforced) schedule to a PRF (partial reinforcement, i.e., the response is reinforced with some probability) schedule they reduce the rate of approaching the asymptote compared to the control group that received reinforcement with the same PRF probability (successive negative contrast effect), (Crespi, 1995; Black, 1968; Cox, 1975). If subjects reinforced on a PRF schedule are switched to a CRF schedule they usually perform at a higher level (rate) than the control group that was exposed only to the CRF schedule (elation effect or positive contrast effect), (Benefield et al., 1974; Maxwell et al., 1976). In its present form, the theory presented here cannot handle positive contrast effects because positive changes in reinforcement conditions cannot be detected (the output of the behavioral inhibition unit can only be positive). However, I illustrate below the operation of the theory in conditions resembling successive and simultaneous negative contrast, an effect reliably obtained in operant studies, (Crespi, 1952; Nevin & Shettleworth, 1966; Bernheim & Williams, 1967; Franchina & Brown, 1971). Among the previous theories showing contrast effects I mention here Grossberg's (1981) theory which explains one type of contrast, behavioral contrast effects, that is a property of operant schedule interactions, which is different from the contrast effect discussed in this paper (Mackintosh, 1974), and Daly and Daly's (1982) model which discusses the same type of negative contrast effects as explained here, i.e., successive and/or simultaneous contrast.

Simulation Results. To test the basic effect, the model is exposed to 14 acquisition trials, where each trial consists of 2000 time units (between trials there is no interval, and therefore the use of the word trial might seem arbitrary; however, I found it useful to divide the training period into trials such that I can better refer to the moment when the change in reinforcement schedule occurs). During each trial animals receive either a small (probability 10%) or a large (probability 100%) reinforcement. The control group receives a small reinforcement for all the 14 trials. The "shifted" group receives a large reinforcement for the first seven trials, and then is switched to the small reinforcement for the remaining seven trials. Figure 7A shows that after the shift occurs, the performance (response strength) gradually becomes higher in the control group. According to the theory, when the negative shift occurs the animal overpredicts the reinforcement. The dynamics of this overprediction is controlled by the long-term and short-term memory for response–reinforcement associations (w_{LM} and w_{SM}). Figure 7B shows that the "shifted" group has a higher reinforcement expectancy than the control group, despite the discontinuity that follows trial 7 that decreases w_{LM} (negative contrast). The difference in the time constants at which w_{SM} and w_{LM} decay as an effect of reinforcement diminution (small reinforcement) contributes to a positive mismatch between the long-term and short-term reinforcement expectancy units ($X_{LE}-X_{SE}>0$). Figure 7C shows that for the "shifted" group the expectancy mismatch is able to trigger the behavioral inhibition unit (X_{BI}) that detects thereby the change in reinforcement contingency [see also eqn (8)]. In this way, X_{BI} inhibits the response [see eqn (1)] observed in the shifted group, a fact that accounts for the negative contrast effect.

The simulation results (Figure 7A) show gradual changes in response strength following the shift in reinforcement magnitude. This result is similar with many negative contrast studies (e.g., Meyer, 1951; Spence, 1956; Bower, 1961; Di Lollo & Beez, 1966) reporting gradual rather than abrupt changes in performance. The gradual changes suggest that the development of the negative contrast effect is driven by learning processes (as hypothesized in the present paper), distinct from motivational variables.

Among the determinants of the negative contrast effects, i.e., magnitude, quality, and delay of reinforcement, I analyze here only the influence of reward magnitude (as expressed by a low reward probability) on the strength of the negative contrast effect. The simulation results are shown in Figure 7D, in which the percentage decrease in response strength, while the reinforcement probability in the low-probability schedule component (small reward) is varied between 0.1 and 0.5, is estimated (the probability of the rich component remains unchanged, i.e., 1). The results are consistent with the findings of Mikulka et al. (1967) who showed that the

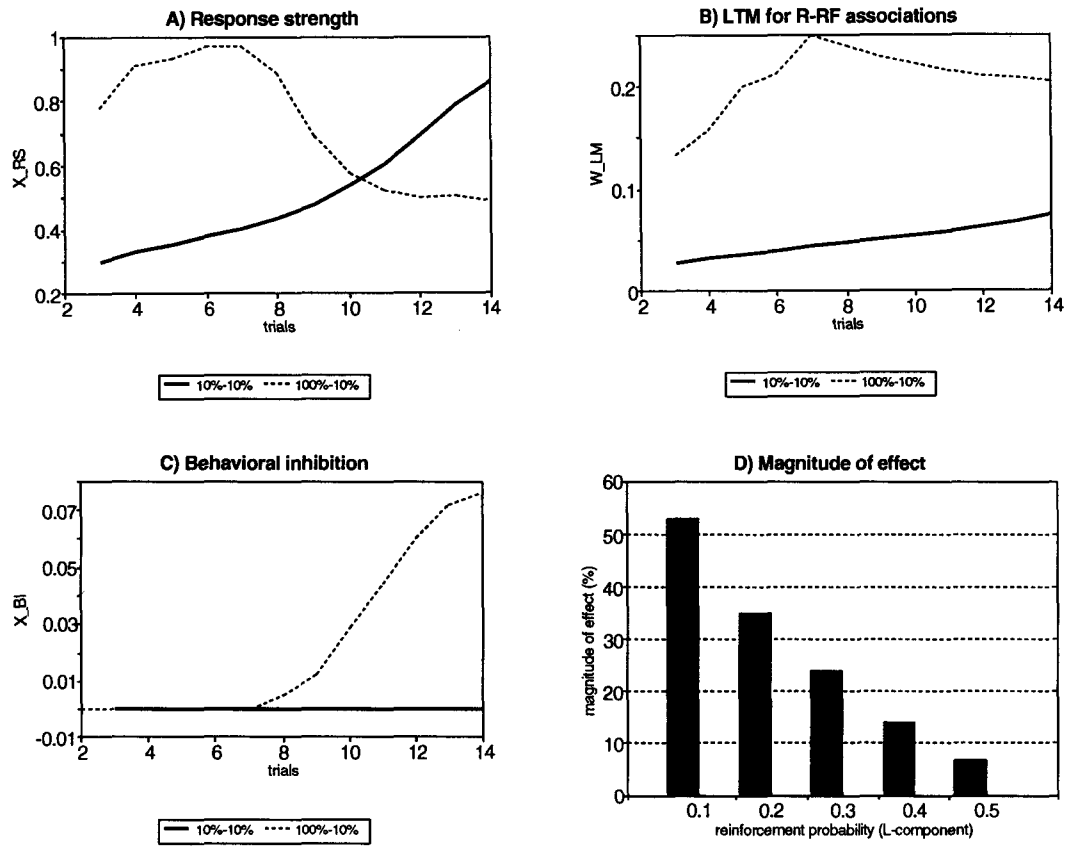

FIGURE 7. Negative contrast effect. (A) The response strength, X_{RS}, in two different situations: (1) PRF schedule (10%) for the whole session as the control condition; (2) continuous reinforcement (100%) followed after trial 7 by PRF schedule (10%) as the shifted condition. Even though the response is much stronger for the CRF segment (as compared to the PRF case), once the schedule becomes intermittent the response strength of the "shifted" group develops at a lower rate than the response strength of the control group. (B) Long-term memory for response–reinforcement associations, w_{LM}: the 'shifted" group has a higher reinforcement expectancy than the control group, despite the discontinuity that follows trial 7. (C) The behavioral inhibition unit, X_{BI}, detects the change in reinforcement contingency by inhibiting the response of the shifted group. (D) Magnitude of effect: percentage decrease in response strength while the reinforcement probability in the *L*-component (small reward) is varied between 0.1 and 0.5.

negative contrast effect can be diminished if the reinforcement discontinuity is made less abrupt.

3.4. Partial Reinforcement Extinction Effect (PREE)

PREE is an effect which has two main aspects: (i) subjects trained to respond to infrequent reinforcement stabilize at a performance value, i.e., rate of responding, generally lower than subjects trained with more frequent reinforcement; (ii) when reinforcement is discontinued (extinction) partially reinforced subjects persist longer in responding than subjects than have been reinforced more frequently, even though they begin extinction responding at a lower rate. The magnitude of the PREE is affected by numerous factors, such as reinforcement probability, pattern of reinforcement, reinforcement delay, reinforcement size, intertribal interval, and length of training. In the present study I will analyze the reinforcement probability as the usual determinant of the PREE. Among the previous dynamic theories dealing with the PREE I mention here Grossberg (1975) and Daly and Daly (1982).

Experimental Data. The basic determinant of the PREE is the probability (or percentage) of reinforcement during acquistion, (Weinstock, 1958; Bacon, 1962; Kacelnik et al.,1987). One of the experimental settings that allows us to analyze this effect is the Kacelnik et al. foraging experiment. In this experiment, starlings chose between two "foraging patches" in which food was delivered according to either a rich or lean probabilistic schedule. There were two comparison groups: in both the lean schedule was 0.08; one group was reinforced with probability 0.25 (rich schedule) and the other one with probability 0.75 (rich schedule). Two main conclusions can be drawn from this experiment: (i) responses reinforced with probability 0.25 are preferred less rapidly than response reinforced with probability 0.75; (ii) after the suppression of reinforcement, responses reinforced with probability 0.25 are more resistant to

extinction than responses reinforced with probability 0.75.

Simulation Results. I have simulated a concurrent probabilistic schedule, relatively similar to the foraging situation described by Kacelnik et al. (1987). The probability of the lean side is held at a fixed level (0.08), whereas the probability of the rich side is varied (1.0, 0.5, and 0.25). As shown in Figure 8A, the preference develops faster when the reinforcement probability on the rich side increases (result already discussed in the *Response selection* section). After 25 000 responses recorded on both sides during acquisition, the reinforcement is extinguished. The proportion of responses on the rich side is calculated for each block of 250 responses. According to Figure 8A, it is clear that responses reinforced with probability 0.25 are more resistant to extinction than responses reinforced with probability 0.5, which are more resistant to extinction than responses reinforced with probability 1.0, even though during acquisition the proportion of responses on the rich side is positively correlated with the reinforcement probability.

The fact that the theory predicts more resistance to extinction with more intermittent schedules can be explained in the following way: the richer the acquisition schedule the higher the level at which both short and long-term memory traces for *R–RF* associations (w_{SM} and w_{LM}) increase. When extinction begins, both w_{SM} and w_{LM} decay, but at different rates; w_{SM} decays at a rate that is higher than w_{LM}'s decay rate, see Figure 8 (panels C and D) and eqns (4) and (5). The dynamics of w_{SM} and w_{LM} drive the increase of X_{BI} via the mismatch between the aggregate long-and short-term reinforcement expectancy units ($X_{LE}-X_{SE}$). Here X_{BI} is 0 during acquisition (see Appendix), but increases during extinction as soon as the difference $X_{LE}-X_{SE}$ becomes greater than 0. Figure 8B shows that the

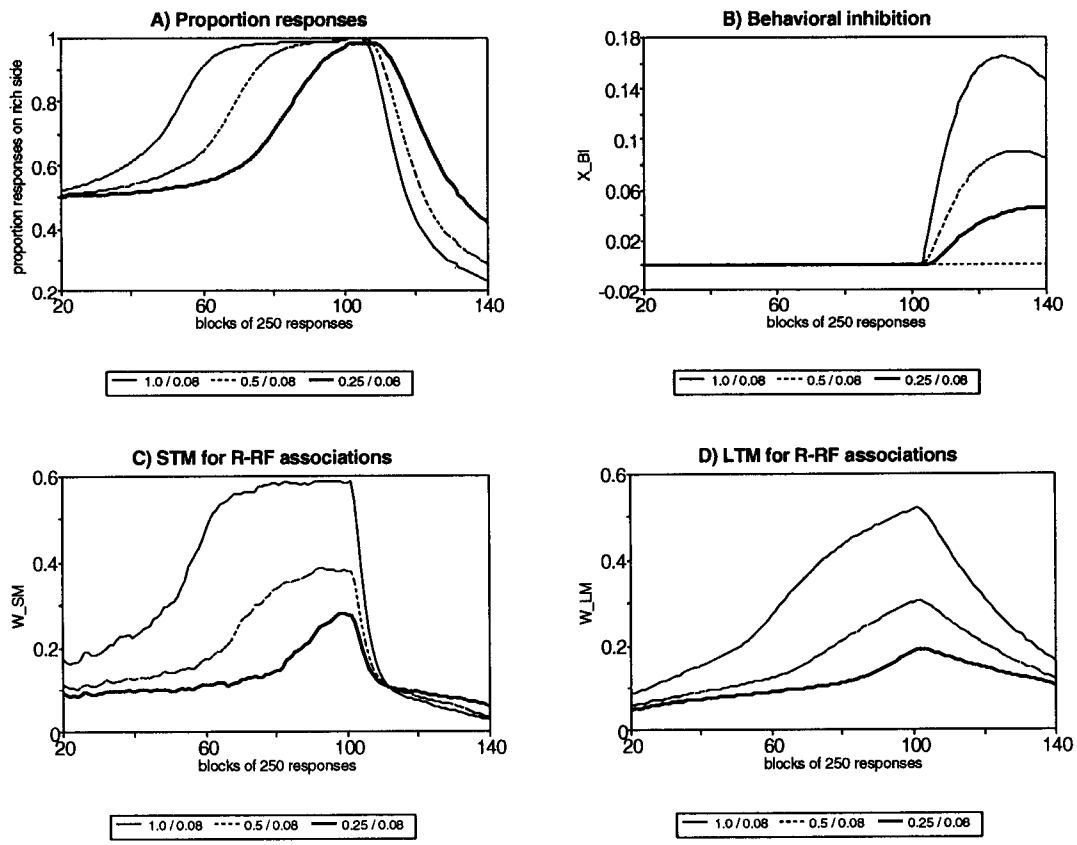

FIGURE 8. Partial reinforcement extinction effect. (A) Proportion responses on rich side as a function of blocks of 250 responses. The acquisition is slower with the decrease in the absolute value of reinforcement probability. Less resistance to extinction with the richer schedule of reinforcement during acquisition. (B) Behavioral inhibition, X_{BI}: the effect of a positive difference $X_{LE}-X_{SE}$ during extinction leads to the increase of X_{BI} to a level proportional to the reinforcement expectancy. (C) Short-term memory for *R–RF* associations; w_{SM}: during acquisition w_{SM} increases to a level proportional to the actual reinforcement probability. After 25 000 responses, the reinforcement is extinguished and w_{SM} decays at a high rate. (D) Long-tem memory for *R–RF* associations, w_{LM}: during acquisition w_{LM} increases to a level proportional to the actual reinforcement probability. The rate of increase is slower and the curve is smoother compared to w_{SM}. After 25 000 responses, the reinforcement is extinguished and w_{LM} decays at a rate that is much slower compared to w_{SM}'s decay rate.

level of the behavioural inhibition unit is proportional to the reinforcement probability during acquisition. Since the rate at which the response strength is inhibited is controlled by the level of the behavioral inhibition unit, there will be more resistance to extinction with more intermittent schedules (Figure 8A). In conclusion, the richer the PRF schedule the higher the level of X_{LE} during acquisition, and thus the higher the level of inhibition (due to X_{BI}) that X_{RS} receives during extinction.

The proposed mechanism by which the difference between the short-term and the long-term rates builds up a mismatch that is used to trigger the behavioral unit during extinction is a robust effect which does not disrupt transient and/or asymptotic behavior. After the asymptote is reached the expectancy mismatch becomes smaller in magnitude, as the difference between the short-term and the long-term memory diminishes (a situation in which the behavioral inhibition is 0). In these conditions, the response relies on the excitatory long-term reinforcement expectancy. However, if the reinforcement probability is suddenly decreased, the equilibrium is broken by the decaying short and long-term memory traces that are able to trigger the behavioral inhibition unit which gradually reduces the response strength (extinction).

3.5. Spontaneous Recovery

Experimental Data.
Changes in performance may occur over an interval of time when the subject is not exposed to reinforcement contingency, and even when the subject is not exposed to the experimental situation at all (effects of inter-session time). One interesting instance of such 'spontaneous" change is spontaneous recovery after extinction. Suppose the animal experiences an acquisition session which is followed by extinction, and then is returned to the experimental chamber. When the new session starts (after some interval since the termination of the prior extinction) the subject's initial pattern of response increases as a function of the inter-session time. If the mean recovery ratio (defined as proportion of responding relative to the total amount of responding during acquisition) is represented as a function of the postextinction interval, two distinct phases can be noticed: (i) during the first days following extinction (usually the first 2 days) recovery reaches a maximum estimated at roughly 40% of the initially acquired response pattern; (ii) after the recovery maximum is reached, the amount of recovery dissipates very slowly in time such that responding can be observed even after 1 week following the end of extinction (Mackintosh, 1974; Robbins 1990).

Spontaneous recovery has not been explained yet, and even though verbal theories have been advanced, the mechanism of spontaneous recovery continues to be a challenge for theorists of conditioning. Most of the proposed verbal theories view recovery as reflecting the central properties of extinction (Pavlov, 1927; Capaldi, 1967, 1971; Rescorla & Wagner, 1972; Mackintosh, 1974). Other theories view recovery as a procedural artifact that has little in common with the basics of extinction (Skinner, 1950; Burstein, 1967). Quantitative models of spontaneous recovery of the operant response are very few. Estes (1955) presented a molar model of spontaneous recovery, which, unfortunately, is unable to explain the mechanism for recovery (the explanations are at a molar level). The CE model (Davis et al., 1993) can only account for a particular form of spontaneous recovery, i.e., regression, that is encountered in choice experiments when, in extinction, there is a reversion to an earlier preference, despite the fact that this alternative is no longer rewarded.

Simulation Results.
To test the operation of the theory in spontaneous recovery I have simulated an acquisition session of 15 000 time units in which the reinforcement is given according to three different probabilistic schedules, with probabilities 0.6, 0.4, and 0.2, followed then by extinction. Figure 9A shows the amount of recovery as a function of the postextinction interval, measured for a reinforcement probability of 0.6. Each point on the curve represents the proportion of responses in 10 000 extinction time units relative to the number of responses recorded during the last 10 000 acquisition time units. The abscissa represents the postextinction interval measured in "days" (1 "day" is equivalent to 30 000 time units). In all two phases are to be distinguished: after the response becomes fully extinguished it recovers within an interval (estimated roughly as 1.5 "days") to about 30% of the initial level of responding, followed by a relatively slower decay that can last for many "days".

Figure 9B shows that, for three different reinforcement probabilities, increasing the time since the offset of extinction can yield more recovery, i.e., the response pattern becomes more persistent (each X_{RS} pulse becomes broader with the time since extinction). The explanation for the dynamics of recovery is the following: in the absence of concurrent alternatives, the response is influenced by two activities [see eqn (1)]: one excitatory, $X_{LE}(w_1 + w_{CLM})(1-X_{RS})$ expressing the influence of long-term reinforcement expectancy, and one inhibitory, $-\alpha_3 X_{BI} X_{RS}$, expressing the inhibitory influence exerted as a consequence of nonreinforcement. The superposition of these two effects generates the profile depicted in Figure 9A.

In the absence of reinforcement the strength of all the connections tends to decrease. This general decay

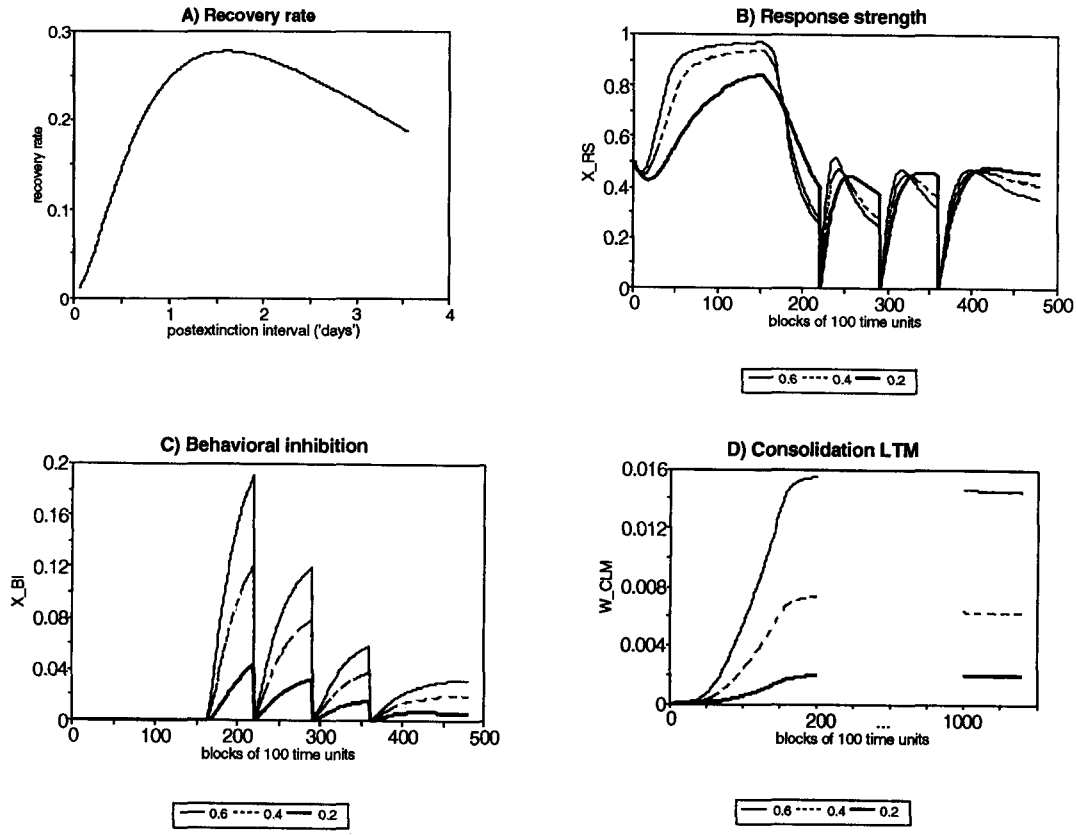

FIGURE 9. Spontaneous recovery. (A) Recovery rate as a function of the postextinction interval. Two phases are to be distinguished: after the response becomes fully extinguished it recovers within an interval (estimated roughly at 1.5 "days") to about 30% of the initial level of responding, followed by a relatively slower decay that can last for many "days". (B) Response strength, X_{RS}, *as a function of time. As the time since the end of extinction elapses the response pattern becomes more persistent. (C) Behavioral inhibition,* X_{BI}, *is negatively correlated with the time since extinction. (D) Consolidation LTM,* w_{CLM}, *gradually increases during acquisition to levels proportional to reinforcement probability. After extinction begins* w_{CLM} *decays slowly.*

process happens at different rates, depending on the functional role of each connection, i.e., w_{RF} and w_{SM} decay fast, w_{LM} decays slower, and w_{CLM} decays even slower. It has been previously shown that due to the decay of the memory units the behavioral inhibition unit increases to maximum during extinction (until the mismatch between the aggregate long and short-term reinforcement expectancy becomes 0) and then decays spontaneously. Figure 9C shows that the behavioral inhibition, measured for each reinforcement probability, decays with the time since extinction and consequently inhibits the response strength unit at lower levels. Concomitant with triggering a positive behavioral inhibition signal, the aggregate reinforcement expectancy also excites the response strength unit. The level of extinction, is controlled by the connection strength $w_1 + w_{CLM}$. Immediately after extinction, the strength of inhibition induced by the behavioral inhibition unit is more effective than the excitatory effect induced by the aggregate reinforcement expectancy (which is modulated by the slowly varying consolidation LTM). As time elapses and the behavioral inhibition decays faster than the consolidation LTM, the response strength is slowly released from inhibition and, due to the excitation from the long-term expectancy it increases to a maximum that coincides with the time when the behavioral inhibition becomes 0. After the behavioral inhibition becomes 0, the response strength decreases slowly (the effect of recovery can last for many days), as it relies only on the slowly decaying consolidation LTM, thus explaining the second portion (after the recovery rate reaches a maximum) of the curve depicted in Figure 9A.

The role of the discriminative stimulus is to trigger the operant behavior after a prolonged absence of reinforced training (during spontaneous recovery). In these conditions, after the strength of all the connections decays to zero, the consolidation LTM is the only variable which preserves information about the history of reinforcement. Whenever the animal is reintroduced in the experimental box the discriminative stimulus switches form 0 to 1 and triggers a burst of responding (in the case of a nonzero consolidation LTM). The responses thus generated read out the consolidation LTM and

increase the response strength at a rate controlled by w_{CLM} (the higher w_{CLM} the higher the recovery rate). Because these "recovery" responses are generated during extinction (when $X_{LE} > X_{SE}$), the behavioral inhibition unit is able to become active and to suppress responding.

4. DISCUSSION

The present study describes a novel theory of operant conditioning in terms of a real-time neural network. The process of conditioning involves the formation of associations between competing operant responses (and environmental stimuli) and the reinforcement. These associations are used to build stimulus and response-specific short- and long-term reinforcement expectancies. The operant response is controlled by the excitatory long-term reinforcement expectancy and by the behavioral inhibition signal triggered by the mismatch between the long- and the short-term reinforcement expectancy. The theory has provisions for historical effects of reinforcement: with extended training the efficacy of response control by the expected reinforcement increases via slow changes in the consolidation long-term memory associations.

4.1. Discussion on Implementation Issues

At least three important issues need to be discussed in relation to the present conditioning model: (i) how are the simulation results influenced by changes in the structure of the model? Is the present configuration a minimal one? (ii) how robust is the model to changes in parameters? What are the most critical parameters? (iii) how well do simulation results fit quantitative aspects of experimental data?

4.1.1. *Model's Sensitivity to Change in Structure.* The present theory builds on six general principles (enumerated in Section 2) that are projected onto the neural network circuit described in Figures 1 and 2. During the testing phase of the model I have tried various alternatives to the present implementation, especially focusing on how far one can go with a reduced set of behavioral principles. In this sense, I will give three examples.

Response-reinforcement associations, w^{R}_{RF}. There are two different ways to refer to these associations. One way, utilized in the present implementation, is to calculate the response associative strength as encoded in a Hebbian connection controlled by the temporal overlap between the response and the reinforcement. Another way to refer to *R–RF* associations is to multiply the stimulus trace at time t with the value of the reinforcement at the same time ($X_{RT}RF$), the result of this operation being further utilized to build short- and long-term memory traces of *R–RF* associations (this method could eliminate w^{R}_{RF}). However, even though more economical in principle, the second possibility did not prove useful for one important reason: if the animal is exposed to a previously reinforced situation it shows a pattern of spontaneous responding despite the fact that no reinforcement is given (spontaneous recovery). In this case the product X_{RF} *RF* would be 0, and the input to the memory units is 0, thus causing the absence of recovery.

Behavioral inhibition, X_{BI}. The operant response receives excitation from the aggregate long-term reinforcement expectancy and it receives inhibition from the behavioral inhibition unit. Another way to implement the control of the operant response could be by utilizing an ensemble of units, behavioral inhibition (X_{BI}) and behavioral excitation (X_{BE}). Here, X_{BI} increases whenever *R–RF* associations are overpredicted, i.e., $X_{LE} - X_{SE} > 0$, and decreases whenever *R–RF* associations are underpredicted, i.e., $X_{LE} - X_{SE} < 0$. Also, X_{BE} increases whenever *R–RF* association are underpredicted, i.e., $X_{SE} - X_{LE} > 0$, and decreases whenever *R–RF* associations are over predicted, i.e., $X_{SE} - X_{LE} < 0$. This implementation has the advantage of offering resources to handle effects of positive changes in reinforcement contingency (such as positive contrast effects, impossible to solve with the actual configuration of the theory). I have already tested this variant of the model, and it shows results comparable to those obtained using the actual implementation.

Consolidation LTM, w_{CLM}. Consolidation LTM is a variable which reflects the history of reinforcement throughout training, and is controlled by the correlation between the aggregate long-term reinforcement expectancy and the operant response. If w_{CLM} is fixed, e.g., 0, the effects of this modification are not observed in situations which involve short durations of acquistion. However, in situations in which the length of training is considerably large, a fixed w_{CLM} has the disadvantage that it does not allow one to discriminate between an "experienced" and a naive animal. I found that a fixed w_{CLM} cannot explain effects such as spontaneous recovery, improvement in performance following discrimination learning, as well as the effect of length of training on the PREE.

4.1.2. *Model Sensitivity to Changes in Parameters.* The Appendix shows the numerical values of the parameters used in the simulations. Each unit in the model has a fixed discharge rate (in the absence of any stimulation) equal to 0.00001. Loosely speaking,

I have implemented four different time-scales: very short (response trace), short (response associative strength, short-term memory for *R–RF* associations), long (long-term memory for *R–RF* associations, interresponse inhibition strength), and very long (consolidation LTM). Correspondingly, I have limited the parameter search to time constants of the following orders: 10^{-1} (very short), 10^{-2} (short), 10^{-4} (long), and 10^{-5} (very long). With all these constraints regarding the parameter space, it was relatively easy to find a configuration of values that generate curves whose profile and relative time course qualitatively match the experimental data. The parameter set indicated in the Appendix probably represents one of the best configurations with respect to graphical appearance. The dynamics of the model are quite robust to perturbations in the parameter space. I found that the most sensitive parameters are those controlling the response strength unit, i.e., the con-nections between X_{BI}–X_{RS}, X_{LE}–X_{RS}, and X_{RS}–X_{RS}.

4.1.3. *Qualitative vs. Quantitative Fits of Experimental Data.*
It is difficult to accommodate within the same framework all the quantitative aspects of the data referenced in this article, mainly for two reasons: the data emerge from experiments with various time-courses, ranging from days (spontaneous recovery) to one session (development of preference), and from experiments involving various species, e.g., pigeons, rats, or starlings. Therefore, the time unit of all the simulations (equal to the numerical integration step using the forward Euler method) was chosen such that the model is able to handle all of the applications under investigation. In this respect, it is not surprising that the time course of the majority of simulations is not identical to the time course of experimental data (the time course is expanded for all the paradigms with the exception of spontaneous recovery where the time course is compressed). Computer simulations (not presented in this article) showed that it is possible to give a better estimate of the real time course of the experiments analyzed here by carefully selecting one integration time step for each paradigm under study.

4.2. Other Issues of Animal Learning

Operant conditioning. One of the most important debates in operant learning is centered around the question *what is hidden behind the statement 'Reinforcement increases the response strength?''* The common view is that, by definition, reinforcement contributes to the increase in the associative strength of certain responses depending on the rate at which those responses have been reinforced. Actually, this interpretation is not always correct. In operant conditioning the only important factor is the *causal* relationship between reinforcement and the response, i.e., reinforcement contingency, not reinforcement or responses considered separately. The theory proposed in this article is able to grasp this important dimension by building the whole model around the notion of response-reinforcement associations, as opposed to the Rescorla–Wagner types of theories which consider the difference between the US and the total reinforcement expectancy as the basis of conditioning. One example that supports this idea is the situation in which free reinforcement (response-independent) decreases the strength of the operant response (Cohen et al., 1993).[3] To explain this effect, let us consider that during the standard training conditions the reinforcement contingency ensures that as reinforcement accumulates the response increases in strength. If no response is emitted, the corresponding response trace decays until it reaches 0. If a reinforcement is offered in these conditions, the value of the multiplication $X_{RF} \cdot RF$ drops to 0, a fact which determines the decrease in response associative strength, w^{R}_{RF} (see Equation 3). If the delivery of response-independent reinforcement goes on, the response associative strength continues to decrease causing the decay of the short and long-term memory traces, a situation which resembles extinction. In these conditions, the mismatch between the long and the short-term expectancy units detects the change in reinforcement contingency and determines the increase in the output of the behavioral inhibition unit which reduces the response strength.

Another example that supports the central role played by response–reinforcement associations in operant conditioning is the effect of the devaluation of reinforcement. If the reinforcement is devalued, for instance by satiation (Adams, 1982; Colwill & Rescorla, 1985), the response is weakened, but it does not disappear entirely. The theory handles this effect in the following way: if the reinforcement is devalued (for instance, *RF* decreases from 1 to 0.1) the shift in the reinforcement value creates conditions for reinforcement overprediction: the association between *R* and *RF* acquires a lower level than before devaluation, a situation detected by the memory units which start decaying. If the reinforcement continues to be delivered at a low level (intensity 0.1), the short-term reinforcement expectancy decreases faster than the long-term reinforcement expectancy. The mismatch between the two expectancy units triggers the behavioral inhibition unit which reduces the strength of the operant response, but the response does not

[3] The classical conditioning analog of this situation is free US given in the absence of the CS.

extinguish, instead it is maintained at a lower rate by the long-term reinforcement expectancy.

Classical conditioning. In order to explore the power of the theory beyond its description of operant learning, I have analyzed the model's behavior in various clasical conditioning paradigms, showing that it would be possible to obtain a unified framework for both types of conditioning. The model correctly describes (1) acquisition of delay and trace conditioning, (2) acquisition with different CS and US durations and intensities, (3) extinction, (4) discrimination acquisition, (5) saving effects, (6) discrimination reversal, (7) generalization, (8) overshadowing. Other paradigms of clasical conditioning such as negative and positive patterning, feature positive discrimination, or conditioned inhibition can only be approached if the model considers *CS–CS* interactions, a refinement left for further applications of the theory.

Avoidance behavior. The present theory can be adapted to incorporate principles that allow to explore acquisition and extinction of the avoidance response. The appetitive response and the avoidance response can be represented by two mutually inhibiting response units, corresponding to appetitive and aversive situations. Two observations are necessary: (i) a positive US (value + 1) corresponds to a positive reinforcement (appetitive), and a negative US (value −1) corresponds to a negative reinforcement (aversive). Both USs are applied to both types of response units; (ii) at the level of the appetitive response unit the short-term US expectancy is an inhibitory unit and the long-term US expectancy is an excitatory unit (as in the present configuration); at the level of the avoidance response unit the short-term US expectancy is an excitatory unit and the long-term US expectancy is an inhibitory unit. The US controls the formation of two types of R–US associations: appetitive (positive associations) and aversive (negative associations), depending on the type of US present in the environment (both types of associations are found at the level of both types of response units). If the negative US is presented in conjunction with a specific warning stimulus, negative associations are formed at the level of both response units. As the aversive conditioning progresses, both avoidance and appetitive response associative strengths become more negative. At the same time, the long and short-term US expectancies continue to increase at different negative levels (more negative for the short-term US expectancy). The negative long-term US expectancy inhibits the appetitive response and excites the avoidance response (via inhibitory connections). The difference X_{LE}–X_{SE} becomes positive for both types of RUs, a situation which triggers a positive $X_{BI}^{appetitive}$ which inhibits the appetitive response and a negative $X_{BI}^{aversive}$ as well as a negative $X_{BI}^{aversive}$ which excite the avoidance response. If the positive US is presented, the situation explained previously reverses, the avoidance response being inhibited and the appetitive response being excited (by its corresponding long-term US expectancy). If the negative US is estinguished the difference X_{LE}–X_{SE} becomes negative for both types of RUs, a situation which triggers a positive $X_{BI}^{aversive}$ (which inhibits the avoidance response) and a negative $X_{BI}^{appetitive}$ (which does not affect the appetitive response). If the positive US is extinguished the difference X_{LE}–X_{SE} becomes positive for both types of RUs, a situation which triggers a positive $X_{BI}^{appetitive}$ (which inhibits the appetitive response) and a negative $X_{BI}^{aversive}$ (which does not affect the avoidance response).

4.3. Comparison with other Learning Theories

As mentioned in the opening section, several theories have been proposed to account for effects of conditioning, either Pavlovian or instrumental. Having defined the principles of the theory and how the model reacts to complex operant conditioning situations, I will now discuss how the theory relates to some of the most representative animal learning models.

4.3.1. *Grossberg* (1972, 1975, 1982). Grossberg proposed that the processing of expected and unexpected events is realized by the interaction between two functionally complementary subsystems. Expected events are processed within an attentional subsystem that establishes the internal representations of responses to expected cues, and unexpected events are processed within an orienting subsystem tht enables the attentional subsystem to adapt to new reinforcement and expectational contingencies. When an unexpected event mismatches an active expectancy within the attentional subsystem, the orienting subsystem both resets the short-tem memory representations within the attentional subsystem and energizes an orienting response. When an expected event matches an active expectancy, the active short-term memory patterns within the attentional subsystem are amplified. At the same time, the amplified (resonant) short-term memory representations induce the formation of adaptive long-term memory changes and inhibit the orienting subsystem. Here two distinct design principles are used to implement Grossberg's ideas: gated dipoles and shunting competitive feedback networks. Gated dipoles consist of parallel on- and off-channels that sustain on-responses to cue onset and off-responses (antagonistic rebounds) to either cue offset or to arousal onset (the antagonistic

rebound is needed to reset the short-term memory patterns within the attentional subsystem). The on- and off-channels are characterized by different transmitter substances (which can act at different rates) that gate the input signals before the on- and off-pathways compete to elicit on- and off-cell short-term memory responses. The inclusion of positive feedback loops turns the gated dipole into a feedback competitive network. The properties of this network are utilized to explain how feedback expectancies interact with feedforward input signals to modulate the active short-term memory representations that influence further the orienting subsystem.

Although Grossberg's models share a number of principles with the present theory, there are differences which should be noted. One important difference refers to the definition of expectancy mismatch. In Grossberg's view the mismatch is computed between the *feedforward input signal* and the *feedback learned expectancy* whereas the present theory considers the *feedforward mismatch between short-term and long-term expectancies*. This difference translates into the goal of each theory: whereas Grossberg's mismatch system detects deviations of the new input signals from the actual expectancy levels, the present theory detects the difference between long- and short-term expectancies of the *association* between stimuli (including the response) and the reinforcement. In this respect, it is not clear how a system that computes the mismatch between the feedforward input and the feedback expectancy can show the spontaneous recovery of a previously extinguished response, given that the post-extinction input stimuli do not contain the reinforcement.

4.3.2. *Daly and Daly* (1982). Daly and Daly assimilated Amsel's concept of frustration (nonreward in the presence of stimuli previously paired with reward arouses an aversive response which becomes classically conditioned to the stimuli present) into a more elaborated model (DMOD) than the Rescorla–Wagner model. DMOD assumes that both smaller and bigger than expected reinforcement are surprising and constitute the basis of conditioning (a measure of surprisingness is given by the difference between the reinforcement and the prediction). DMOD also assumes that 'courage' to approach a negative goal event can be conditioned and is called counterconditioning. Behavior is determined by the total V-value that is assumed to be equal to the summation of three V-values: approach, aversive, and counterconditioning.

Although DMOD and the present theory share an important idea, i.e., learning is driven by the mismatch between expected reinforcement and the current events, the main difference lies in the definition of the event. DMOD considers as event the actual value of the US (either appetitive or aversive), whereas the present theory considers as events the association between responses or discriminative stimuli and the reinforcement, an important distinction that originates from the idea of operant conditioning (not the reinforcement, but its association with the stimuli present is important). In addition, DMOD can only account for trial-by-trial changes, not for real-time changes in behavior. This is one of the reasons why DMOD cannot deal with effects of intersession time, for instance spontaneous recovery.

4.3.3. *Klopf* (1988). Klopf (1988) proposed a neuronal model for classical conditioning in which it is suggested that *changes* in stimuli and *changes* in responses should be correlated instead of correlating simple stimuli and responses (drive–reinforcement, or D–R model). Klopf defines neuronal drives as signal levels and reinforcements as changes in signal levels. A novel idea derived from these definitions is the model's sensitivity to onset and offset of stimuli, consistent with the idea of Mowrer (1960) that the onsets of both CSs and USs are used as reinforcements. The D–R model is a variant of the time-derivative models for reinforcement learning, similar to the Sutton and Barto (1981) model (S–B model) in the use of the time derivative of reinforcement prediction, but different from the S–B model in the use of changes in signal levels as opposed to signal traces.

The D–R model and the present theory share the idea of changes in reinforcement prediction as a basic mechanism for conditioning. However, whereas the D–R model considers as relevant the changes which occur at consecutive time steps, the present theory considers as relevant the difference between predicted reinforcement (long-term expectancy) and experienced events (short-term expectancy). Another difference between the two models is that whereas the D–R model considers changes in stimulus levels associated with changes in response levels (ΔS–ΔR associations), the present theory considers changes in the strength of the *association* between stimulus levels and the response, $\Delta(S$–$R)$ associations, also different from Thorndike's (1911) stimulus-response (S–R) associations.

4.3.4 *Sutton and Barto* (1990). Sutton and Barto (1990) have developed a learning algorithm, mainly applied to phenomena from classical conditioning, which uses the difference between reinforcement predictions, computed at successive time steps (discrete-time analog of the time derivative of the prediction), as the basis of conditioning (TD model). The model also makes use of STM stimulus eligibility traces that help the formation of stimulus–reinforcement associations. In essence, the TD model

hypothesizes that the goal of learning is the prediction at each point in time of the imminence-weighted sum of future reinforcement.

The TD model and the present theory share several assumptions, such as the inclusion of eligibility traces or the use of the time-difference between successive reinforcement predictions in the learning rule. Although it is true that, in principle, both theories have in common the idea of mismatch between different temporal estimates of the same variable, i.e., reinforcement expectancy (or prediction), the actual implementation of reinforcement expectancy is very different. The TD model uses the difference between successive predictions of *reinforcement* whereas the present theory uses the difference between long- and short-term predictions of *associations* between stimuli (including the response) and the reinforcement. In other words, TD model proposes to predict the reinforcement, whereas the present theory proposes to predict the *contingent* reinforcement. This important distinction is reflected into the data that the TD model would have difficulty to account for. For instance, if the CS is turned off and free reinforcement is offered in this condition, the TD model would predict that the stimulus associative strength remains constant, a fact that contradicts the contingency effect (response strength or CR decreases with the accumulation of free reinforcement).

4.3.5. *Schmajuk and DiCarlo* (1992). Schmajuk and DiCarlo (1992) used a "generalized" delta rule (changes in the synaptic weights between two neural populations are performed by minimizing the squared value of the difference between the output of the population and the actual US) to train a layer of hidden units that "configure" simple CSs (S–D model). The input layer is activated by conditioned stimuli and the context stimulus and forms direct associations with a first output layer. Another set of associations are formed with the hidden layer, and this layer in turn forms new associations with a second output layer. A backpropagation procedure, which differs from the standard method in that the error signal (expressed as the mismatch between the US and the aggregate prediction of the US) instead of including the derivative of the activation function of the hidden units simply contains their activation function, is utilized to train the hidden-unit layer. The output of the hidden layer encodes the "configural stimuli", in fact the internal representation of the CSs.

Even though the S–D model has been applied to explain data from classical conditioning, there are principles which it shares with the present operant conditioning theory. Both models consider the existence of STM stimulus traces that help the formation of S–RF associations, although the S–D model does not present resources for response–reinforcement associations. Both models assume that information is processed based on the mismatch between current events and past events. However, the S–D model is a goal-seeking supervised learning system (it minimizes the error between the actual reinforcement and the value of the prediction generated by all the CSs), whereas in the present theory the difference between the aggregate long- and short-term reinforcement expectancy along with the aggregate long-term reinforcement expectancy are the variables that control the response.

4.3.6. *Davis et al.* (1993). According to Davis et al. (1993), the core of recurrent choice is a competitive learning process in which ratios between the number of responses and the number of reinforcements at the level of each response unit, computed from the whole history of training, are compared according to a winner-take-all strategy (CE model). The theory presented in this article shares with the CE model the idea of sensitivity to the history of reinforcement (consolidation LTM). Another common feature is the use of a competitive learning rule (although different in mathematical implementation) to express response selection. Nonetheless, the assumptions of the CE model are not sufficient to explain effects which require provisions for real time or rates of occurrence, such as spontaneous recovery or the PREE. Furthermore, the CE model does not incorporate any behavioral principle that is sensitive to changes in reinforcement contingency (for instance, contrast effects).

4.4. Concluding Remarks

In the present article I have attempted to solve problems related to the role of reinforcement, response, and discriminative stimulus in operant conditioning by proposing a neural network theory that describes a behavioral mechanism of conditioning. The model explains a wide range of operant conditioning phenomena (some of the most representative are presented in Section 3), and many implications of the theory are considered for both classical conditioning and for the development of the avoidance response.

Among the ideas advanced in this article, four principles are crucial in the conception of the present theory: (1) It is hypothesized that the *goal* of operant conditioning is to accurately predict the contingent reinforcement by defining the reinforcement expectancy as the aggregate prediction of R–RF associations. This idea is contrasted with the ideas advanced by other theories of conditioning which asume that the goal of learning is to accurately predict the actual or the future reinforcement levels, e.g., Klopf (1988), Sutton and Barto (1990), Schmajuk and DiCarlo

(1992). In this respect, I propose a Δ(S–R) behavioral theory (the learning mechanism is driven by changes in the association between stimuli and the response). (2) Different time scales used in the detection of novel events: short-term memory—reduced time scale, long term memory—intermediate time scale, consolidation LTM—large time scale. (3) The aggregate reinforcement expectancy controls the rate of increase of the operant response. (4) The response is controlled by the behavioral inhibition unit which integrates the mismatch between expected and experienced events [resembles the functions of Gray's (1971) behavioral inhibition system and Amsel's (1962) frustration theory].

Finally, it is worth noticing that the present theory is not a complete theory of operant conditioning. Among the major phenomena left aside by the theory are: (1) Learning reinforced patterns of switching (no provision for response–response associations). (2) Fixed-ratio schedules of reinforcement (effects of postreinforcement pause). (3) Timing (especially effects of fixed-interval reinforcement schedules). However, these problems of the present theory are viewed as targets for futurue developments of the model rather than critical flaws in the framework.

REFERENCES

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, **34B**, 77–98.

Amsel, A. (1962). Frustrative nonreward in partial reinforcement and discrimination learning. *Psychological Review*, **69**, 306–328.

Bacon, W. E. (1962). Partial-reinforcement extinction effect following different amounts of training. *Journal of Comparative and Physiological Psychology*, **55**, 998–1003.

Bailey, J. T., & Mazur, J. E. (1990). Choice behavior in transition: development of preference for the higher probability of reinforcement. *Journal of the Experimental Analysis of Behavior*, **53**, 409–422.

Benefield, R., Oscos, A., & Ehrefreund, D. (1974). Role of frustration in successive positive contrast effect. *Journal of Comparative and Physiological Psychology*, **86**, 648–651.

Bernheim, J. W., & Williams, D. R. (1967). Time-dependent contrast effects in a multiple schedule of food reinforcement. *Journal of the Experimental Analysis of Behavior*, **10**, 243–249.

Black, R. W. (1968). Shifts in magnitude of reward and contrast effects in instrumental conditioning. *Psychological Review*, **75**, 114–126.

Bower, G. H. (1961). A contrast effect in differential conditioning. *Journal of Experimental Psychology*, **62**, 196–199.

Burstein, K. R. (1967). Spontaneous recovery: A (Hullian) noninhibition interpretation. *Psychonomic Science*, **7**, 389–390.

Capaldi, E. J. (1967). Sequential versus nonsequential variables in partial delay of reward. *Journal of Experimental Psychology*, **74**, 161–166.

Capaldi, E. J. (1971). Memory and learning: a sequential viewpoint. In W. K. Honig & P.H.R. James (Eds.). *Animal memory* (pp. 111–154). New York: Academic Press.

Capaldi, E. J., & Minkoff, R. (1967). Reward schedule effects at a relatively long intertrial interval. *Psychonomic Science*, **9**, 169–170.

Chung, S. H., & Herrnstein, R. J. (1967). Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior*, **10**, 67–74.

Cohen, S. L., Riley, D. S., & Weigle, P. A. (1993). Tests of behavior momentum in simple and multiple schedules with rats and pigeons. *Journal of the Experimental Analysis of Behavior*, **60**, 255–291.

Collwill, R. M., & Rescorla, R. A. (1985). Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *Journal of Experimental Psychology: Animal Behavior Processes*, **11**, 520–536.

Cox, W. M. (1975). A review of recent incentive contrast studies involving discrete trial procedures. *The Psychological Record*, **25**, 373–393.

Crespi, L. P. (1952). Quantitative variation of incentive and performance in the white rat. *American Journal of Psychology*, **55**, 467–517.

Daly, H. B., & Daly, J. T. (1982). A mathematical model of reward and aversive nonreward: its application in over 30 appetitive learning situations. *Journal of Experimental Psychology: General*, **111**, 441–480.

Davis, D. G. S., Staddon, J. E. R., Machado, A., & Palmer, R. (1993). The process of recurrent choice. *Psychological Review*, **100**, 320–341.

Di Lollo, V., & Beez, V. (1966). Negative contrast effect as a function of magnitude of reward decrement. *Psychonomic Science*, **5**, 99–100.

Dragoi, V., & Staddon, J. E. R. (1993). A competitive neural network model for the process of recurrent choice. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, & A. S. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School* (pp. 65–73). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, **62**, 145–154.

Franchina, J. J., & Brown, T. S. (1971). Reward magnitude shifts effects in rats with hippocampal lesions. *Journal of Comparative Physiological Psychology*, **76**, 365–370.

Gray, J. A. (1971). *The psychology of fear and stress.* London: Weidenfeld and Nicholson.

Gray, J. A. (1982). *The neuropsychology of anxiety: an enquiry into the functions of the septo-hippocampal system.* New York: Oxford University Press.

Grossberg, S. (1972). A neural theory of punishment and avoidance, II: Quantitative theory. *Mathematical Biosciences*, **15**, 39–67.

Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology*, **18**, 263–325.

Grossberg, S. (1981). Psychophysiological substrates of schedule interactions and behavioral contrast. *SIAM–AMS Proceedings*, **13**, 157–186.

Grossberg, S. (1982). Processing of expected and unexpected events during conditioning and attention: a psychophysiological theory. *Psychological Review*, **89**, 529–572.

Gutman, A. (1977). Positive contrast, negative induction, and inhibitory stimulus control in rat. *Journal of the Experimental Analysis of Behavior*, **27**, 219–233.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, **56**, 51–65.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of Experimental Analysis and Behavior*, **4**, 267–272.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, **13**, 243–266.

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Kacelnik, A., Krebs, J. R. & Ens, B. (1987). Foraging in a changing

environment: an experiment with starlings (sturnus vulgaris). In M. L. Commons, A. Kacelnik & S. J. Shettleworth (Eds.), *Quantitative analyses of behavior VI: foraging* (pp. 63–87). Hillsdale, NJ: Laurence Erlbaum.

Killeen, P. (1968). on the measurement of reinforcement frequency in the study of preference. *Journal of the Experimental Analysis of Behavior*, **11**, 263–269.

Killeen, P. (1970). Preference for fixed-interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior*, **14**, 127–131.

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiology*, **16**, 85–125.

Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. New York: John Wiley.

Mackintosh, N. J. (1974). *The psychology of animal learning*. New York: Academic Press.

Maxwell, F. R., Calef, R. S., Murray, D. W., Shephard, D. C. & Norville, R. A. (1976). Positive and negative successive contrast effects following multiple shifts in reward magnitude under high drive and immediate reinforcement. *Animal Learning and Behavior*, **4**, 480–484.

Mazur, J. E. (1992). Choice behavior in transition: development of preference with ratio and interval schedules. *Journal of Experimental Psychology: Animal Behavior Processes*, **18**, 364–378.

Mazur, J. E. (1995). Development of preference and spontaneous recovery in choice behavior with concurrent variable-interval schedules. *Animal Learning and Behavior*, **23**, 93–103.

McEwen, D. (1972). The effects of terminal-link fixed-interval and variable-interval schedules on responding under concurrent chained schedules. *Journal of the Experimental Analysis of Behavior*, **18**, 253–261.

McSweeney, F. K., Roll, J. M. & Weatherly, J. N. (1994). Within-session changes in responding during several simple schedules. *Journal of the Experimental Analysis of Behavior*, **62**, 109–132.

Meyer, D. R. (1951). The effects of differential rewards on discrimination reversal learning by monkeys. *Journal of Experimental Psychology*, **41**, 268–274.

Mikulka, P. J., Lehr, R., and Pavlik, V. B. (1967). Effect of reinforcement schedules on reward shifts. *Journal of Experimental Psychology*, **74**, 57–61.

Minsky, M. L. (1963). Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, **49**, 8–30, 1961. Reprinted in E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 406–450). New York: MacGraw-Hill.

Mowrer, O. H. (1960). *Learning theory and behavior*. New York: Wiley (Krieger Edition, 1973).

Nevin, J. A. (1988). Behavioral momentum and the partial reinforcement effect. *Psychological Bulletin*, **103**, 44–56.

Nevin, J. A., & Shettleworth, S. J. (1966). An analysis of contrast effects in multiple schedules. *Journal of the Experimental Analysis of Behaviour*, **9**, 305–315.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, **87**, 532–552.

Rescorla, R. A., & Wanger, 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.

Reynolds, B. (1961). Behavioral contrast. *Journal of Experimental Analysis of Behavior*, **4**, 57–71.

Robbins, S. J. (1990). Mechanisms underlying spontaneous recovery in autoshaping, *Journal of Experimental Psychology: Animal Behavior Processes*, **16**, 235–249.

Rudy, J. W. (1971). Sequential variables as determiners of the rat's discrimination of reinforcement events: effects of extinction performance. *Journal of Comparative Physiological Psychology*, **77**, 476–481.

Rudy, J. W., & Sutherland, R. J. (1989). The hippocampal formation is necessary for rats to learn and remember configural discriminations. *Behavioral Brain Research*, **34**, 97–109.

Samuel, A. L. (1963). Some studies in machine learning using the game of checkers. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill. (Reprinted from *IBM Journal on Research & Development*, 1959, **3**, 210–229.)

Schmajuk, N. A. (1995). *Animal learning and cognition: A neural network approach*. Cambridge: Cambridge University Press (in press).

Schmajuk, N. A., & DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review*, **99**, 268–305.

Schwartz, B., & Gamzu, E. (1977). Pavlovian control of operant behavior: an analysis of autoshaping and its implication for operant conditioning. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior*. Englewood Cliffs, NJ: Prentice Hall.

Skinner, B. F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, **57**, 193–216.

Sokolov, E. N. (1960). Neuronal models and the orienting reflex. In M. A. B. Brazier (Ed.), *The central nervous system and behavior, 3rd Conference* (pp. 187–276). New York: Josiah Macy Jr Foundation.

Spence, K. W. (1956). *Behavior theory and conditioning*. New Haven, CN: Yale University Press.

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

Staddon, J. E. R., & Hinson, J. M. (1978). Behavioral competition: a mechanism for schedule interactions. *Science*, **202**, 432–434.

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* (pp. 279–393). Hillsdale, N.J.: Lawrence Erlbaum.

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

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. W. Moore, (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks*. Cambridge, MA: MIT Press.

Swanson, L. W. (1978). The anatomical organization of septo-hippocampal projections. In K. Elliot & J. Whelan (Eds.), *Functions of the septo-hippocampal system* (pp. 25–43). Ciba Foundation Symposium 58 (New Series).

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

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York: Appleton-Century-Crofts.

Vinogradova, O. S. (1975). Functional organization of the limbic system in the process of registration of information: facts and hypotheses. In R. L. Isaacson & K. H. Pribram (Eds.), *The hippocampus, v. 2, Neurophysiology and behavior* (pp. 1–70). New York: Plenum Press.

Weinstock, S. (1958). Acquisition and extinction of a partially reinforced running response at a 24-hour intertrial interval. *Journal of Experimental Psychology*, **47**, 151–158.

APPENDIX

All the units described in the text vary between 0 and 1. The numerical value of X_{BI} as used in simulations is obtained by taking $max(0, X_{BI})$ computed at each time step. All the computer simulations have been performed equating the integration step (0.15 for all the simulated paradigms) with a formal time unit (1 s). A *RF* of magnitude 1 is applied according to all the simulated reinforcement schedules. Parameter values used in all simulations are $\alpha_1 = \alpha_8 = 1 \times 10^{-5}$ $\alpha_9 = 3 \times 10^{-5}$, $\alpha_2 = 23 \times 10^{-4}$, $\alpha_7 = 5.2 \times 10^{-4}$, $\alpha_5 = \alpha_6 = \alpha_{10} = 0.01$, $\alpha_3 = 0.08$, $w^R = w^S = 0.1$, $w_1 = 0.15$, $\alpha_4 = 0.5$.