Learning & Behavior
2010, 38 (1), 50-67
doi:10.3758/LB.38.1.50

Two-factor theory, the actor–critic model, and conditioned avoidance

Tiago V. Maia
Carnegie Mellon University, Pittsburgh, Pennsylvania

Two-factor theory (Mowrer, 1947, 1951, 1956) remains one of the most influential theories of avoidance, but it is at odds with empirical findings that demonstrate sustained avoidance responding in situations in which the theory predicts that the response should extinguish. This article shows that the well-known actor–critic model seamlessly addresses the problems with two-factor theory, while simultaneously being consistent with the core ideas that underlie that theory. More specifically, the article shows that (1) the actor–critic model bears striking similarities to two-factor theory and explains all of the empirical phenomena that two-factor theory explains, in much the same way, and (2) there are subtle but important differences between the actor–critic model and two-factor theory, which result in the actor–critic model predicting the persistence of avoidance responses that is found empirically.

Avoidance behavior is behavior that results in the omission of an aversive event that would otherwise occur. Several theories of avoidance have been proposed (e.g., Bolles, 1970, 1972a, 1972b; Herrnstein, 1969; Mowrer, 1947, 1951, 1956, 1960; Seligman & Johnston, 1973). Of these, Mowrer's (1947, 1951, 1956) two-factor theory was, and arguably still is, the most influential (see, e.g., Domjan, 2003; Levis & Brewer, 2001; McAllister & McAllister, 1995). Two-factor theory had its origins in Hull's (1943) suggestion that all conditioned responses are established via drive reduction. For Hull, *drives* were states such as hunger, the reduction of which acts as a reward. Mowrer (1947, 1951, 1956) proposed that fear is also a drive and that reductions in fear are therefore also rewarding. According to Mowrer (1947, 1951, 1956), however, fear itself is learned not by drive reduction but by Pavlovian pairing of conditioned stimuli (CSs) and aversive unconditioned stimuli (USs). The name *two-factor theory* highlights this use of two learning principles.

Figure 1 illustrates one of the most common experimental paradigms in the study of avoidance: the so-called *discriminated avoidance procedure*, which will be the focus of this article. In this paradigm, a trial starts with the presentation of a warning stimulus (often called a CS). An aversive US is scheduled to occur a certain time after the onset of the CS, but animals can avoid the US if they perform a predetermined response (e.g., crossing to the other side of a shuttle box) before that time. Typically, that response also terminates the CS. Mowrer (1947, 1951, 1956) attributed the learning of the avoidance response in this paradigm to the termination of the CS rather than to the actual avoidance of the US. He proposed that the CS comes to elicit fear because early in training, before the avoidance response is learned, the CS is often followed by the US. When the animal subsequently performs a response that terminates the CS, fear is reduced, thereby strengthening the response.

The present article shows that the actor–critic model (Barto, 1995; Barto, Sutton, & Anderson, 1983), which has previously been used to account for a wealth of behavioral and neural findings in conditioning (Daw, 2003; Houk, Adams, & Barto, 1995; Joel, Niv, & Ruppin, 2002; Maia, 2009; O'Doherty et al., 2004; Suri, Bargas, & Arbib, 2001; Takahashi, Schoenbaum, & Niv, 2008; Z. M. Williams & Eskandar, 2006), provides a natural computational implementation of the main ideas of two-factor theory. The article further shows that there are subtle but important differences between the actor–critic and two-factor theory that allow the actor–critic to explain several empirical findings in discriminated avoidance that are at odds with two-factor theory.

The ideas, model, and simulations in this article were first presented in my doctoral dissertation (Maia, 2007). A related model was later proposed by Moutoussis, Bentall, Williams, and Dayan (2008). Their model, however, was based on advantage learning (Baird, 1993; Dayan & Balleine, 2002), whereas the model presented here is based on the standard actor–critic (Barto, 1995; Barto et al., 1983; Sutton & Barto, 1998). In addition, their main emphasis was on the effects of dopaminergic manipulations, whereas the emphasis here is on the relation of the actor–critic to two-factor theory. Other computational models of avoidance learning that bear close resemblance to two-factor theory have been proposed (Grossberg, 1972; Johnson, Li, Li, & Klopf, 2002; Schmajuk & Zanutto, 1997). Of these, the model with the closest resemblances to the

T. V. Maia, tmaia@columbia.edu

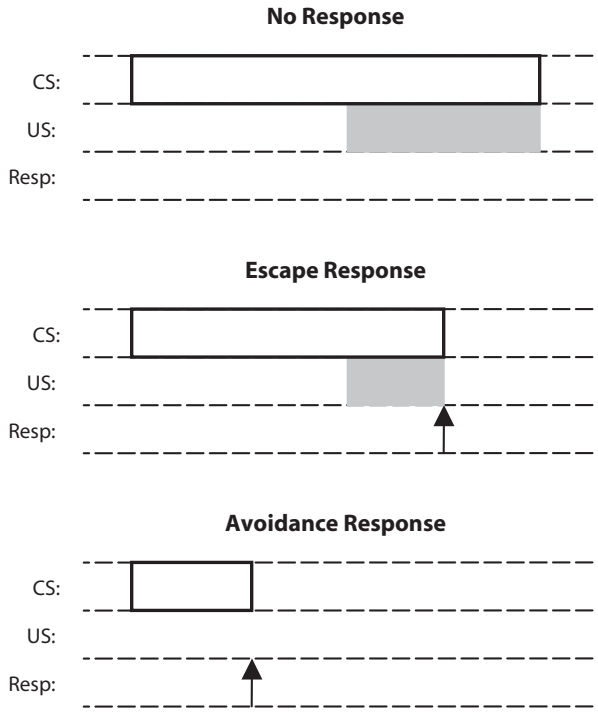

Figure 1. The discriminated avoidance paradigm. Top: No response. At the start of each trial, a warning stimulus (often called a CS) is presented and remains on. If the animal does not perform a predetermined response (e.g., crossing to the other side of a shuttle box), after a predetermined period of time an aversive US (e.g., a shock) comes on. If the animal still does not perform the response, the CS and US stay on for a predetermined period of time, after which they both terminate. Middle: Escape response. If the animal performs the response after the US comes on, both the CS and the US are immediately terminated. Such trials are called escape trials, because by performing the response the animal escapes the US. Bottom: Avoidance response. If the animal responds after the CS comes on but before the US starts, the CS is terminated and the US is not presented. Such trials are called avoidance trials, because by performing the response the animal avoids the US.

one presented here is that by Johnson et al. (2002), which was based on Klopf, Morgan, and Weaver's (1993) associative control process (ACP) framework. The ACP framework is closely related to two-factor theory (Klopf et al., 1993) and to the actor–critic (Johnson et al., 2002), but the use of standard reinforcement learning machinery in the present article has the advantage of linking more directly to the extensive work on reinforcement learning models of conditioning and to the well-developed computational theory of reinforcement learning. The models of Grossberg (1972) and Schmajuk and Zanutto (1997), while capturing several empirical findings and being sources of important insights, were more exclusively motivated by the behavioral findings, without independent computational motivation, and without connecting to the known neural bases of conditioning.

THE ACTOR–CRITIC MODEL

The building blocks of any reinforcement learning problem are *states*, *actions*, and *reinforcements*. States represent the current state of the external or internal world. For example, being on the left side of a shuttle box may be represented as one state and being on the right side as another. When the agent is in a given state, it selects an action (e.g., jumping over the barrier on a shuttle box) from among those available in that state. Partly as a result of that action, the agent may then transition to a new state (e.g., the other side of the shuttle box). Different states may be associated with different reinforcements (e.g., shock on one side of the shuttle box and no reinforcement on the other). I will represent the reinforcement associated with state s as a scalar $R(s)$.

The value $V(s)$ of state s is the expected sum of future reinforcements when the agent starts in state s. If we represent the reinforcement received at time t as r_t and the current time as τ, then

$$V(s) = E\left\{\sum_{t=\tau}^{\infty} \gamma^{t-\tau} r_t \mid s_\tau = s\right\},$$

where γ is a discount factor that makes reinforcements that happen farther in the future count less than more proximate reinforcements, s_τ represents the state at time τ, and E is the expected value operator.

A powerful method for learning the values of states is the method of temporal differences (Sutton, 1988). Suppose that the agent is in state s and performs some action a that takes it to state s', where it receives reinforcement $R(s')$. Before the action, the agent's estimate of the value of s is $V(s)$. After the action, the agent's new estimate of the value of s is the sum of the reinforcement that it actually received and the value of s' (where the latter is discounted by γ). Formally, the agent's new estimate of the value of s is $R(s') + \gamma V(s')$. At first sight, the agent could then simply update its prior estimate of $V(s)$ to $R(s') + \gamma V(s')$. However, because the action selection, the transitions between states, and the reinforcements may all be stochastic, it is best to treat $R(s') + \gamma V(s')$ as a sample and to update $V(s)$ just a little in the direction of that sample. This is done using the *prediction error* δ, which is the difference between the estimated value of $V(s)$ given by this one sample and the prior estimate of $V(s)$:

$$\delta = R(s') + \gamma V(s') - V(s). \quad (1)$$

The agent then updates its estimate of $V(s)$ so as to reduce this error in the future:

$$V(s) \leftarrow V(s) + \alpha\delta, \quad (2)$$

where α is a learning-rate parameter.

Prediction errors indicate how things turned out relative to what was expected: Positive prediction errors indicate that things turned out better than expected, and negative prediction errors indicate that things turned out worse than expected. Prediction errors can therefore also be used to learn which actions are advantageous. Let $p(s,a)$ represent the strength of the S–R association between state (or stimulus) s and action (or response) a. If the agent performs action a in state s and things turn out better than it expected, it should strengthen the association between s and a—that is, increase $p(s,a)$. If things turn out worse

than expected, $p(s,a)$ should be decreased. This can be accomplished with the following equation:

$$p(s,a) \leftarrow p(s,a) + \beta\delta, \tag{3}$$

where β is a learning-rate parameter.

Prediction errors can therefore be used to simultaneously learn the values of states and the preferences for actions. The actor–critic is a connectionist architecture that implements these ideas. The architecture consists of three main components: the state, the critic, and the actor (Figure 2). The units in the state layer represent the current state. All of the simulations in this article use a localist representation in which there is one unit per state. To represent a given state, the corresponding unit is on and all other units are off. I will refer to the unit that represents state s simply as state unit s.

The critic learns the values of states and calculates prediction errors. It consists of two units: the value prediction unit (represented by a circle in Figure 2), whose output is $V(s)$, and the prediction-error calculation unit (represented by a square in Figure 2), whose output is δ. The prediction-error calculation unit calculates the prediction error using Equation 1. With a localist representation for the state and with linear activation functions, as were used here, the synaptic weight from state unit s to the value prediction unit represents $V(s)$ (see, e.g., Maia, 2009). The weights between the state units and the value prediction unit are therefore learned according to Equation 2.

The actor learns the preferences for actions and selects actions accordingly. It consists of a layer of action units. With a localist representation for states and actions and with linear activation functions, as were used here, the synaptic weight from state unit s to action unit a represents $p(s,a)$ (see, e.g., Maia, 2009). These weights are therefore learned according to Equation 3. When the system is in state s, each action unit becomes activated with activity $p(s,a)$. To select an action, these activations are first transformed into a probability distribution by application of a softmax rule. The softmax is a generalization of the maximum operator: Instead of simply selecting the most active action unit, the softmax gives each unit a probability of being selected that preserves the rank order of the activations (Bridle, 1990). In other words, actions whose action units are more activated have a higher probability of being selected. Specifically, the probability $\pi(s,a)$ of taking action a in state s is given by

$$\pi(s,a) = \frac{e^{p(s,a)/\tau}}{\sum_{b \in A(s)} e^{p(s,b)/\tau}}, \tag{4}$$

where b ranges over all actions available in state s [represented by $A(s)$] and τ is a *temperature* parameter. Application of Equation 4 to all actions available in state s gives a probability distribution over those actions. An action is then chosen stochastically from that distribution. The temperature τ controls the degree of randomness in action selection. When τ is large, the actions are more equiprobable, so action selection is more random; when τ is small, the difference in probability between actions with different values of $p(s,a)$ is sharpened, so action selection is more deterministic.

RELATION OF THE ACTOR–CRITIC TO TWO-FACTOR THEORY

Two-factor theory proposed two learning mechanisms: classical conditioning, based on the Pavlovian pairing of CSs and USs, and instrumental conditioning, based on drive reduction. These two mechanisms correspond closely to the critic and the actor, respectively.

We saw that the critic uses temporal-difference learning to learn the values of states—that is, to learn to predict the reinforcement that is signaled by the different states.

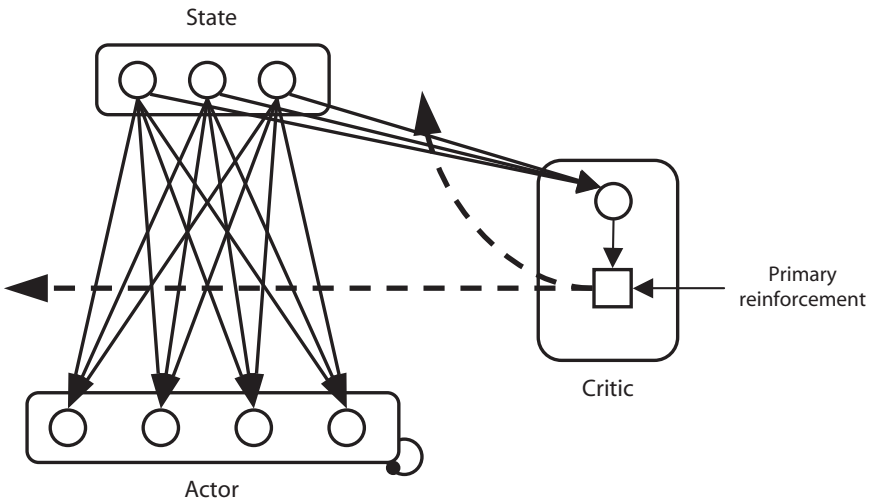

Figure 2. The actor–critic architecture. See the text for details. From "Reinforcement Learning, Conditioning, and the Brain: Successes and Challenges," by T. V. Maia, 2009, *Cognitive, Affective, & Behavioral Neuroscience*, *9*, p. 348. Copyright 2009 by the Psychonomic Society. Reprinted with permission.

Virtually all quantitative theories of classical conditioning suggest that classical conditioning corresponds precisely to learning to predict future reinforcements (the USs) on the basis of the CSs (e.g., Dayan, Kakade, & Montague, 2000; Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Sutton & Barto, 1990). It is therefore natural to use temporal-difference learning (i.e., the critic) to model classical conditioning (e.g., Schultz, 1998; Sutton & Barto, 1990). In fear conditioning, the idea is that fear corresponds to the prediction of an aversive outcome. If we represent aversive outcomes with a negative scalar, fear will correspond to a negative value. For example, suppose that in state s_1 there is an aversive US, such as shock, so $R(s_1) < 0$. Now, suppose that state s_2 is a CS that predicts this US. As we will see below, through pairings of the CS and the US, $V(s_2)$ will become negative. We can take that negative value to correspond to fear.

In short, the critic essentially implements the classical conditioning component of two-factor theory. The other factor in two-factor theory is instrumental conditioning via drive reduction. Two-factor theory suggests that actions are reinforced when they result in an *immediate* reduction in a drive; for example, avoidance actions are reinforced because, by terminating the CS, they result in an immediate reduction in fear. The same thing happens with the actor, as we will now see.

An action in the actor is strengthened (weakened) if it results in an *immediate* positive (negative) prediction error (see Equation 3). The actor therefore implements Thorndike's (1911) law of effect, but with an important difference: Instead of strengthening or weakening actions on the basis of the primary reward or punishment that immediately follows them, it strengthens or weakens actions on the basis of the prediction error that immediately follows them. Now, going back to Equation 1, we can see that the prediction error corresponds to the sum of the primary reinforcement, $R(s')$, with a difference between values, $\gamma V(s') - V(s)$. In the context of fear and avoidance conditioning, the values reflect fear, which in two-factor theory is considered an acquired drive. The difference $\gamma V(s') - V(s)$ therefore reflects the change in that acquired drive. In the context of fear and avoidance conditioning (and ignoring the discount factor γ, for simplicity), if this difference is positive it means that we went from a negative state s to a less negative state s'—in other words, from a state that elicited more fear to a state that elicited less fear. Thus, a reduction in fear results in a positive prediction error, which strengthens the response, just as proposed by two-factor theory.

In summary, there are deep similarities between the actor–critic and two-factor theory. The critic learns the values of states, which correspond to acquired drives, and the actor learns which actions to perform on the basis of the immediate prediction error, which partly reflects changes in drive strength. In the context of fear and avoidance conditioning, the values of states learned by the critic correspond to fear; a reduction in fear leads to a positive prediction error, which reinforces the preceding response in the actor.

SIMULATIONS USING THE ACTOR–CRITIC MODEL

Actor–Critic Model for One-Way Avoidance Experiments

For simplicity, all of the simulations in this article are focused on one-way avoidance, which is an instance of the discriminated avoidance procedure described above. Its particular characteristics are that at the beginning of each trial the animal is always put in the same side of a shuttle box, and the avoidance or escape response consists of crossing to the other side of the box. Figure 3 shows a Markov decision process (MDP) representation of one-way avoidance (see also Smith, Li, Becker, & Kapur, 2004).

Although intuitive, the representation in Figure 3 suffers from an important limitation: It does not represent the passage of time within each state. This is problematic because some experimental data deal with response latencies. Another limitation of the MDP in Figure 3 is that its only terminal state is the one labeled "Safe side; CS off; US off." Thus, a trial must continue until the animal crosses to the safe side. In many experiments, however, if the animal does not cross to the safe side after a certain time of being shocked, the trial is terminated. Implementing this also requires representing the passage of time within a state.

A common approach to keeping track of time is to use a tapped delay line (Daw, Courville, & Touretzky, 2006; Daw & Touretzky, 2002; Johnson et al., 2002; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Smith et al., 2004; Sutton & Barto, 1990). The idea is to "unfold" a state into multiple substates that represent the passage of time within the original state. Applying this technique to the MDP in Figure 3 gives the MDP in Figure 4. The MDP in Figure 4 also includes a new state ("End"), which represents the end of the trial if the animal has failed to cross to the safe side after being shocked for a prespecified period of time.

The actor–critic corresponding to the MDP in Figure 4 is shown in Figure 5. Note that each state has been split into n substates. (Different states do not necessarily have to be split into the same number of substates; I use n for all for simplicity.) All of the simulations in this article use the actor–critic architecture from Figure 5, with $n = 4$.

Simulations

I conducted two simulations. Simulation 1 consisted of 60 acquisition trials. Simulation 2 consisted of 15 acquisition trials followed by 510 extinction trials. In one of the early experiments on avoidance, Solomon, Kamin, and Wynne (1953) ran one of their dogs for 490 extinction trials and found no signs of extinction. The use of a very large number of extinction trials in Simulation 2 is intended to show that the model also does not exhibit any signs of extinction, even with more than 490 extinction trials.

Results

This section presents the results of the two simulations. Rather than being organized by simulation, however, the

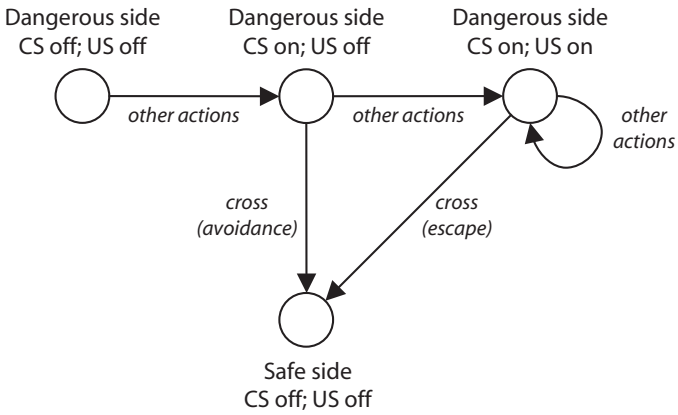

Figure 3. Markov decision process representation of the one-way avoidance paradigm. At the beginning of each trial, the animal is put on the dangerous side (the side in which it can get shocked). Often, at this stage the passage to the other side of the shuttle box is blocked by a closed gate (e.g., Solomon & Wynne, 1953), so the animal cannot cross. After a prespecified interval, regardless of what the animal does, the CS comes on. At this point, there are two possibilities: Either the animal crosses to the safe side (which would make this an avoidance trial) or it does not. In the latter case, after a prespecified interval, the US (typically a shock) begins. Again, there are two possibilities: Either the animal crosses to the safe side (which would make this an escape trial) or it does not. In the latter case, both the CS and the US stay on. Remaining in the state in which the US is on leads to continuing punishment; the animal will therefore want to escape from that state as soon as possible. Regardless of when the animal crosses to the safe side, the gate is typically closed immediately thereafter (e.g., Solomon & Wynne, 1953), so the animal cannot return to the dangerous side. Note that "other actions" means all actions except crossing to the other side of the shuttle box. The reinforcement $R(s)$ associated with state s is written inside the circle that represents that state.

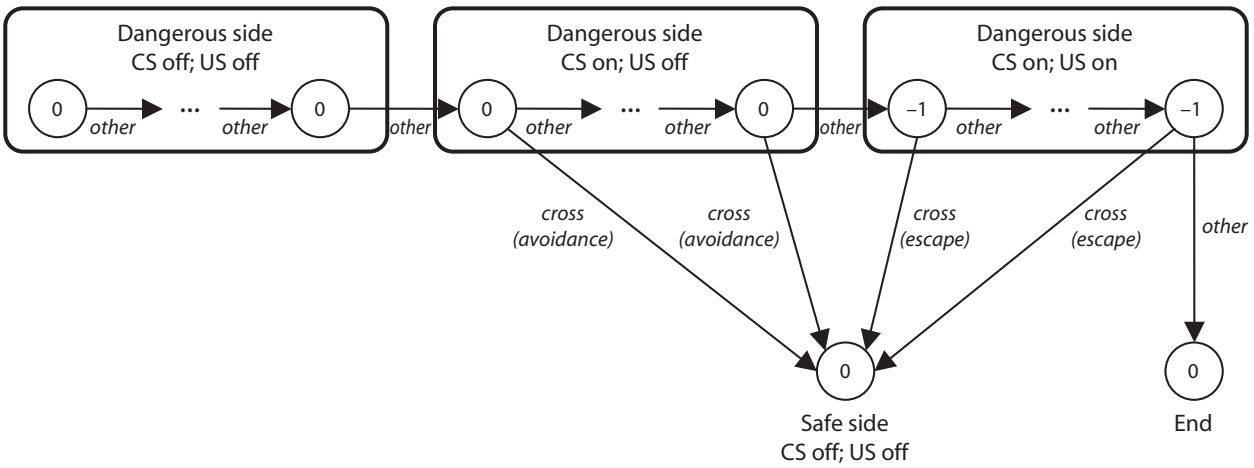

Figure 4. Markov decision process representation of the one-way avoidance paradigm, with substates to represent the time spent in each state. Both the state labeled "Safe side; CS off; US off" and the state labeled "End" are terminal states. The label "other" is shorthand for "other actions."

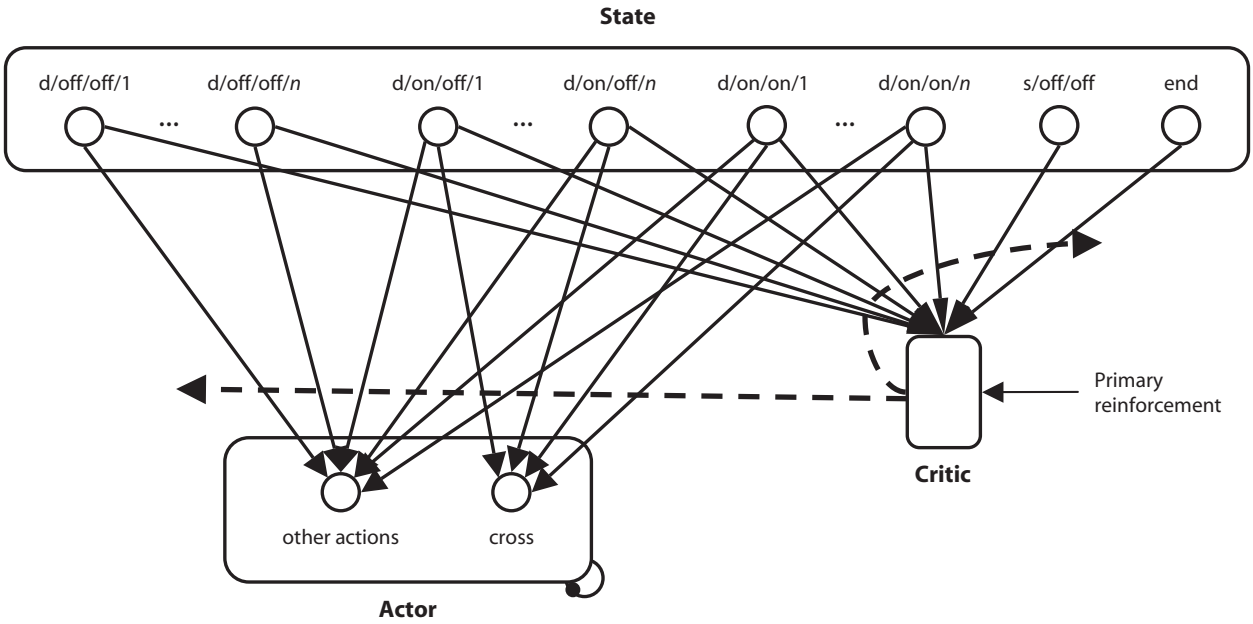

Figure 5. Actor–critic architecture for the one-way avoidance paradigm, with substates to represent the time spent in each state. The names of the states (except for the state "end") consist of three or four components separated by slashes, according to the following conventions: The first component represents the location (d, dangerous side; s, safe side); the second represents whether the CS was on or off; the third represents whether the US was on or off; and the fourth, when present, represents the substate number. Thus, for example, d/on/off/1 represents the first substate of the state that corresponds to the dangerous side with the CS on and the US off. For simplicity, only one unit is shown to represent other actions, but in the simulations 10 such units were present to represent the fact that there were multiple alternative actions. The link between the d/off/off substates and the "cross" action unit is absent because the gate is closed before the CS comes on, so the animal cannot cross in those substates.

presentation is organized by empirical finding. The findings that can be explained by two-factor theory are presented before those that cannot.

Findings Consistent With Two-Factor Theory

Finding 1: Avoidance responses are learned. In the early trials of avoidance conditioning, the animal often gets shocked, because it has not yet learned to avoid the shock. According to two-factor theory, this results in fear of the shock. The same thing occurs in the model. When, in the early trials, the model receives shocks, the states in the top row of Figure 4 come to predict an aversive outcome; that is, they acquire a negative value V. This is illustrated in Figure 6, which shows the values of the states (left column) and the probability of crossing (right column) at several time points in Simulation 1. For now, note only that the states that represent the CS being on (d/on/off/1 through d/on/off/4) acquire a negative value early in training (e.g., after 10 trials); this corresponds to fear of the CS. We will return to the other aspects of this figure below.

According to two-factor theory, once fear of the CS is in place, a response that terminates the CS reduces fear and is therefore reinforced. The same thing occurs in the model. Note in Figure 6 that whereas the d/on/off states acquire a negative value early in training, the safe state always maintains a value of 0. When the model is in one of the d/on/off states and performs the crossing response, it therefore moves from a state with a negative value to a state with a value of 0; this results in a positive prediction error, which strengthens the crossing response. To see the similarity to two-factor theory, note that the positive prediction error occurs because the model goes from a state that evokes fear (a state with a negative value) to a state that does not (a state with a value of 0). Escape responses are learned by a similar mechanism: The d/on/on states have a negative value, so when the model crosses to the safe state from one of the d/on/on states, there is a positive prediction error, which reinforces the response.

Finding 2: Escapes occur before avoidances. Early in training, animals perform mostly escape responses, whereas later in training, they nearly exclusively perform avoidance responses (Beninger, Mason, Phillips, & Fibiger, 1980; Mowrer, 1960; Solomon & Wynne, 1953). This is consistent with two-factor theory. According to two-factor theory, for the avoidance response to be learned, the CS must first elicit fear (so that the termination of the CS is reinforcing). Fear of the CS must itself be learned, so in the early trials there is little or no reinforcement for avoidance responses. Escape responses, in contrast, terminate an aversive US, which is a primary drive, so they are reinforced right from the start of the experiment. The same thing occurs in the model. For the avoidance response to be learned, the states representing the CS must first acquire a negative value. In the model, escapes are therefore also more prevalent in the beginning, whereas avoidances become predominant later in training (Figure 7).

Finding 3: During avoidance training, fear first increases and then decreases. Fear of the CS typically in-

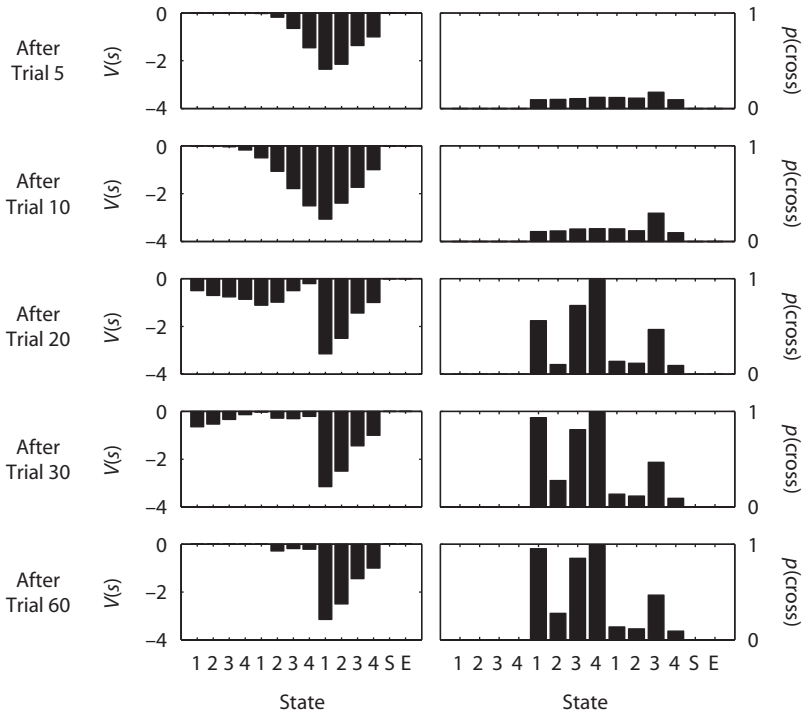

Figure 6. Values of states and probabilities of performing the crossing response over the course of acquisition in Simulation 1. Recall that each of the larger states d/off/off, d/on/off/, and d/on/on is split into four states. States are represented in the *x*-axis, following the same order in which they are presented in Figure 5: d/off/off/1 through d/off/off/4 (labeled 1–4), d/on/off/1 through d/on/off/4 (the second set labeled 1–4), d/on/on/1 through d/on/on/4 (the third set labeled 1–4), s/off/off (labeled "S"), and end (labeled "E"). The probabilities of performing the crossing response are obtained using Equation 4.

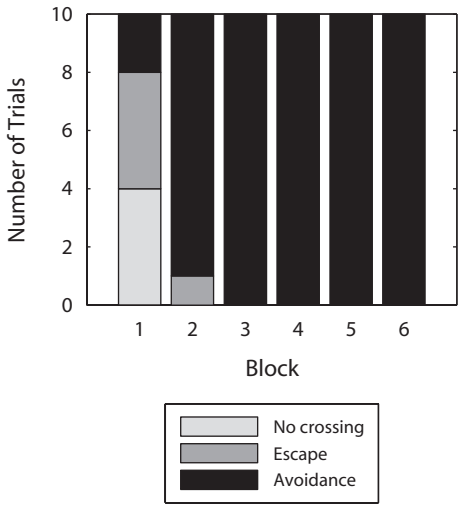

Figure 7. Learning in the model. The graph shows the number of failures to respond ("No crossing"), escape responses, and avoidance responses, per blocks of 10 trials, in Simulation 1. Early in training, failures to respond and escapes predominate; as training proceeds, avoidance responses become more predominant. Starting on the third block, the model always successfully avoids the shock.

creases early in training and then decreases with more extended training (e.g., Brady, 1965; Brady & Harris, 1977; Coover, Ursin, & Levine, 1973; Kamin, Brimer, & Black, 1963; Solomon et al., 1953; Solomon & Wynne, 1953). This is consistent with two-factor theory. Early in training, before the avoidance response is well established, the CS is often followed by the US; this results in fear of the CS increasing. As the avoidance response becomes well established, the CS is rarely, if ever, followed by the US; this results in extinction of the fear. The explanation of the actor–critic is very similar, as we will now see.

To model fear of the CS, I summed the value V of all d/on/off states. To obtain a positive rather than negative measure of fear, I then used the negative of that result. Fear F of the CS was therefore determined using the following equation:

$$F = -\sum_{i=1}^{n} V(\text{d/on/off}/i), \tag{5}$$

where n is the number of d/on/off states ($n = 4$ in all simulations).

Figure 8 shows that in Simulation 1 fear increases early in training and then decreases with more extensive training. The explanation for this pattern in the model is similar to the explanation provided by two-factor theory. In Simulation 1, Trial 11 is still an escape trial, but after that, the model successfully avoids on every trial. Fear therefore decreases from Trial 12 onward.

The fact that fear should decrease in successful avoidance trials is intuitive, but how does that occur in the model? Suppose that the model is in one of the d/on/off states (say, d/on/off/i) and performs an avoidance. If d/on/off/i has a negative value, the avoidance takes the model from a state with a negative value to a state with a value of 0 (the safe state); this results in a positive prediction error, which increases the value of state d/on/off/i (i.e., decreases its absolute value), thereby reducing fear.

Finding 4: Fear does not completely extinguish even with many consecutive avoidances. Despite the widely replicated finding of a decrease in fear with extensive avoidance training, as Mineka (1979) emphasized, "fear has never been demonstrated to extinguish completely . . . even with extensive training" (p. 993). According to two-factor theorists, this is because when the animal performs an avoidance response, only the fear of that part of the CS that the animal is actually exposed to extinguishes (Levis & Brewer, 2001). If, as is found empirically (Solomon & Wynne, 1953), the animal performs the avoidance response with short latencies, the fear to the later part of the CS does not extinguish. The techniques used to assess fear of the CS during avoidance conditioning (e.g., conditioned suppression; Estes & Skinner, 1941) typically present the CS for its entire duration, so they measure fear of the entire CS. According to two-factor theory, the fear measured by these techniques reflects fear of the later, unextinguished part of the CS. The model's explanation is the same.

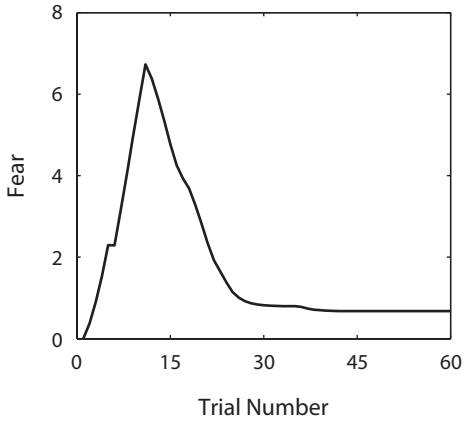

Figure 8. Fear of the CS after each trial in Simulation 1.

Note in Figure 8 that fear of the CS does not return to baseline. To demonstrate that fear in the model does not fully extinguish even with many more consecutive avoidances, I ran an additional simulation with 700 acquisition trials. Starting on Trial 12, the model successfully avoided on every trial. Fear decreased substantially until around Trial 30, then fairly slowly until around Trial 175, then only marginally until around Trial 350, and then remained approximately constant, never returning to baseline, despite the continuing consecutive avoidances.

To understand why fear does not fully extinguish in the model, it is helpful to go back to Figure 6. The left column of Figure 6 shows the value (fear) for each state.[1] The right column shows the probability of performing the avoidance response in each state. As expected, the probability of performing the avoidance response increases with training. Now, consider the situation by Trial 30. The probability of performing the avoidance response in state d/on/off/1 is nearly 1; thus, the model will perform the avoidance response as soon as it reaches this state on virtually every trial. Note, however, that by Trial 30, fear of the subsequent d/on/off states (d/on/off/2 through d/on/off/4) has *not* fully extinguished. Given that the model will perform the avoidance response on the d/on/off/1 state in virtually every trial, there will be little opportunity for further extinction of the fear in states d/on/off/2 through d/on/off/4. This is the reason that by Trial 60 fear for those states has diminished very little. Furthermore, by Trial 60 the probability of performing the avoidance response in state d/on/off/1 is even higher, so there will be even less opportunity for fear of the states d/on/off/2 through d/on/off/4 to extinguish. Thus, fear of those states will tend to persist.

Finding 5: Fear reemerges when the avoidance response is blocked and when the animal responds with a longer latency. We saw above that with extended training, fear of the CS is greatly reduced. However, if the response is blocked, animals exhibit increased fear (e.g., Solomon et al., 1953). Two-factor theory's explanation of this finding is the same as the explanation for why fear does not fully extinguish: When the response is prevented, the animal is exposed to the later, unextinguished parts of the CS, which elicit fear. The model's explanation is the

same. Consider, for example, the situation in the last row of Figure 6. Fear of all of the d/off/off states and of d/on/off/1 has already fully extinguished, and the probability of crossing in state d/on/off/1 is very close to 1. Left to its own devices, the model will then avoid on state d/on/off/1 in virtually every trial, with no concomitant fear. If, however, one blocks the response, the model is exposed to states d/on/off/2 through d/on/off/4, which elicit fear. The same ideas explain why animals (Solomon & Wynne, 1954) and the model exhibit increased fear when they respond with longer-than-usual latencies.

Findings Inconsistent With Two-Factor Theory

Finding 6: Avoidance responses are extremely resistant to extinction. According to two-factor theory, when the avoidance response is well learned and the animal avoids on every trial, fear of the CS should extinguish. Termination of the CS would then no longer be reinforcing, so the avoidance response should also extinguish. If at that point the animal was under an extinction schedule, it would not be shocked again, and the response should remain extinguished. If the animal was instead still under an acquisition schedule, it would get shocked again; this would produce fear conditioning, which would support learning of the avoidance response, and the entire cycle would be repeated. Two-factor theory therefore predicts that under an extinction schedule the avoidance response should extinguish, and under an extended acquisition schedule there should be cycles of response learning and extinction. Neither of these predictions corresponds to the empirical findings. Instead, once the avoidance response is well learned, animals often continue to perform it in every trial, during both acquisition and extinction (Levis, 1966; Levis, Bouska, Eron, & McIlhon, 1970; Levis & Boyd, 1979; Logan, 1951; Malloy & Levis, 1988; McAllister, McAllister, Scoles, & Hampton, 1986; Seligman & Campbell, 1965; Solomon et al., 1953; Solomon & Wynne, 1953; Wahlsten & Cole, 1972; R. W. Williams & Levis, 1991). Furthermore, avoidance responding persists, and even gets stronger, after fear of the CS is drastically reduced or even nearly extinguished (Hodgson & Rachman, 1974; Mineka, 1979; Rachman, 1976; Rachman & Hodgson, 1974; Riccio & Silvestri, 1973). This has been found during schedules involving extended acquisition (e.g., Cook, Mineka, & Trumble, 1987; Kamin et al., 1963; Mineka & Gino, 1980; Neuenschwander, Fabrigoule, & Mackintosh, 1987; Starr & Mineka, 1977) or extinction following acquisition (e.g., Solomon et al., 1953). This persistence of the avoidance response even after fear is greatly reduced is considered one of the main problems for two-factor theory.

We saw above that even though fear is greatly reduced after several consecutive avoidances (Finding 3), it does not fully extinguish (Finding 4). It could therefore be argued that whatever little fear remains is sufficient to maintain the avoidance response. However, as was discussed above, the residual fear that is measured by tests such as conditioned suppression likely reflects fear of the later part of the CS, not of the part of the CS that typically precedes the avoidance response. In fact, this is the prediction of two-factor theory itself. But in that case, one would expect the avoidance response to extinguish, possibly with a gradual lengthening of the avoidance responses preceding such extinction. To see this, suppose that at a certain point the animal is responding at approximately t sec after the onset of the CS. After several such responses, fear of the first t sec of the CS would extinguish, and responding at t sec would no longer be reinforcing. The t-sec response would then extinguish. Later responses, however, would still be reinforced because fear of the CS after t sec would not have extinguished. The animal might therefore start to respond a little later (say, at approximately $t + \Delta t$ sec). However, that would result in extinction of fear of the first $t + \Delta t$ sec of the CS, which would result in extinction of the response at $t + \Delta t$ sec and a shift to a later response. Eventually, this would result in extinction of the response.

The actor–critic can explain the resistance of avoidance responding to extinction, regardless of whether the schedule does or does not switch to extinction trials. In the aforementioned simulation with 700 acquisition trials, the model exhibited no cycles of avoidance learning and extinction. Once the avoidance response was well established (by Trial 12), it was performed in every trial. The same thing occurred in Simulation 2, which consisted of 15 acquisition trials followed by 510 extinction trials. Again, the avoidance response was well established by Trial 12, and it was performed on every subsequent trial, including throughout the 510 extinction trials.[2] The fact that the model performs similarly in these two simulations is not surprising. The model performs the avoidance response on every trial starting on Trial 12, so it makes no difference whether the training regime changes to an extinction schedule after that trial.

Furthermore, in the model the avoidance response persists even with little or no fear. In Simulation 2, for example, once the avoidance response was well established, it continued to be performed in every single trial, despite a marked reduction in fear (Figure 9). It could be argued that perhaps the residual fear in Figure 9 is sufficient to continue to reinforce the response. However, that fear is fear of the later parts of the CS; this is shown more clearly

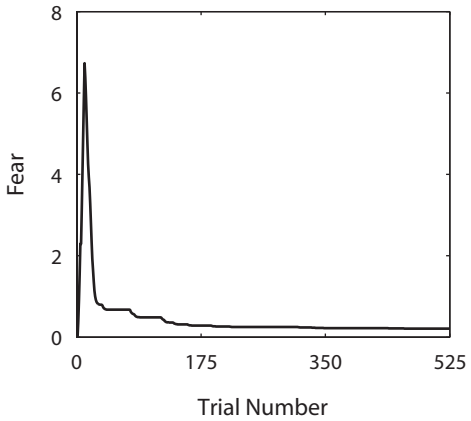

Figure 9. Fear of the CS in Simulation 2.

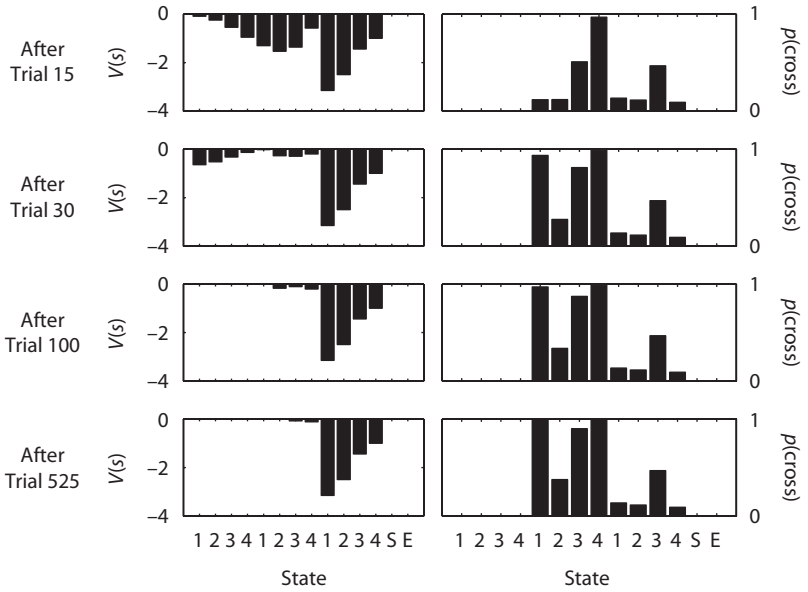

Figure 10. Values of states and probabilities of crossing in Simulation 2.

in Figure 10. Note in particular that by Trial 100, fear of d/on/off/1 and earlier states has already fully extinguished and the probability of crossing in d/on/off/1 is very close to 1. The model will therefore avoid on virtually every trial on state d/on/off/1, which does not elicit fear, so there will be no opportunity for fear reduction to continue to reinforce the response. Although by Trial 100 fear of d/on/off/1 and earlier states has already fully extinguished, 425 trials later the probability of performing the response has not decreased. The persistence of the avoidance response in the model is therefore *not* due to continuing reinforcement via fear reduction; in the model, unlike in two-factor theory, extinction of fear does not result in extinction of the response. But why?

The answer follows from Equation 3, together with the definition of prediction error. Suppose that fear of state s has already extinguished [i.e., $V(s) = 0$]. Now suppose that the model makes an avoidance response in that state. That will lead it to the safe state, which also has a value of 0. The prediction error is therefore 0, which, according to Equation 3, implies that the strength of the response does not change.

This is a key point, so it is worth rephrasing. Equation 3 shows that for the strength of an action to decrease in the model, a negative prediction error is required. However, there is never a negative prediction error when the avoidance response is performed, regardless of whether the experiment consists only of acquisition trials or of a sufficient number of acquisition trials followed by extinction trials. Initially, there is classical conditioning of the CS to the shock, so the states that represent the CS acquire a negative value. Then, when the model performs the response and is not shocked, there is a positive prediction error, which reinforces the response. Once the avoidance response becomes well established, the consecutive avoidances make the value of the states that represent the CS come to 0, because the CS is no longer predictive of the US. The CS therefore comes to predict a reinforcement of 0, and the actual and predicted future reinforcements are also 0 whenever the avoidance response is performed, because the model transitions to the safe state, from where it never receives any shock. The prediction error is therefore 0, so the response is not weakened.

Finding 7: Avoidance latencies decrease throughout acquisition and even extinction. Avoidance latencies decrease throughout acquisition, becoming much shorter than necessary to avoid the shock (Beninger et al., 1980; Solomon & Wynne, 1953). For example, in the experiment of Solomon and Wynne (1953), after 50 trials of training the mean avoidance latency was 2 sec, even though the shock only occurred at 10 sec. Avoidance latencies also continue to decrease during extinction trials (Solomon et al., 1953). This should come as no surprise, since we have already noted that experiments with only an extended acquisition phase and experiments in which an extended acquisition phase is followed by an extinction phase are indistinguishable for animals.

The finding of decreasing latencies throughout extended acquisition and in extinction is problematic for two-factor theory. As we saw above, fear during later phases of acquisition and during extinction is significantly reduced; a reduction in fear is therefore accompanied by an apparent increase in the strength of the response, which is the opposite of what two-factor theory would predict. This, however, is exactly what the actor–critic predicts.

Figure 11 shows the avoidance latencies in Simulation 2. Note that the latencies continue to decrease during extinction. Furthermore, such decrease occurs in parallel with a decrease in fear (Figure 9). The same thing occurs during extended acquisition in Simulation 1 (not shown because the graph is exactly the same as that for Simulation 2). To understand these results, we need to understand

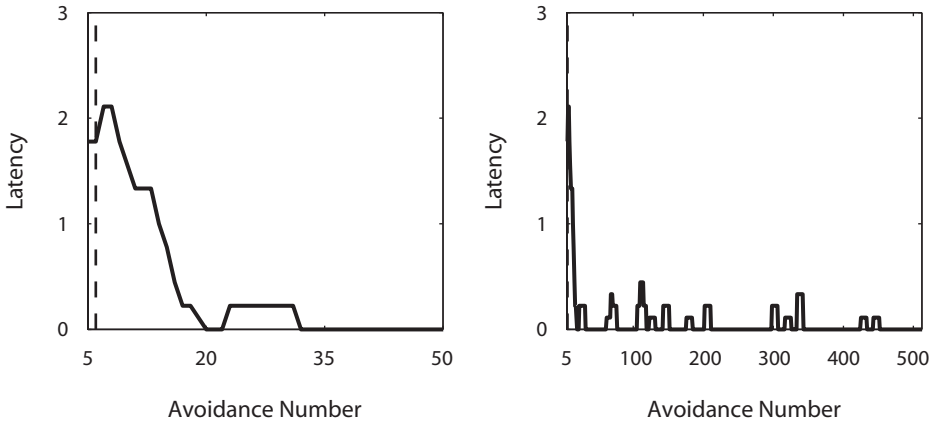

Figure 11. The avoidance latencies in Simulation 2 continue to decrease during extinction. For clarity, the left panel shows only the first 50 avoidances. The right panel shows all avoidances. The vertical axis represents the latency of the avoidance response. In the model, the latency is given by the substate of the d/on/off state in which the crossing occurred. Latencies of 0, 1, 2, and 3 correspond to avoidance responses in states d/on/off/1, d/on/off/2, d/on/off/3, and d/on/off/4, respectively. The horizontal axis represents the number of the avoidance response; for example, 7 represents the seventh avoidance response (not Trial 7). The figure shows a moving average of the avoidance latencies, with a span of nine. Thus, the value plotted at each point *i* is an average of the avoidance latencies from *i*−4 through *i*+4. The moving average is not defined for the first and last four points in the data set. The end of the acquisition phase is marked by a dashed vertical line. There were six avoidances during the acquisition phase in this simulation, so the acquisition phase is over after the sixth avoidance. The avoidance latencies continue to decrease markedly after the end of acquisition. The occasional increases in latency are due to randomness in action selection in the model. The fact that such increases show up as plateaus rather than as single points is an artifact of the smoothing used to construct the figure. The increase in latency in points 7 and 8 is due to this randomness and is not related to the end of the acquisition phase; it also occurs in Simulation 1, which continues the acquisition phase for several more trials. Even though starting in about Trial 35 most avoidances have a latency of 0, there is occasionally an avoidance with longer latency. Given the smoothing imposed by the moving average procedure used to construct the figure, this shows up as a plateau. These plateaus tend to become less frequent (and smaller) as extinction proceeds; thus, with more extinction trials, the avoidance response becomes even more likely to occur with a latency of 0. If one averaged across animals, as is often done in the experimental literature, this would result in a more gradual decrease in latencies.

both the mechanism underlying the reduction in avoidance latencies in the model and why avoidance latencies continue to decrease even as fear markedly decreases. We will start by considering the first question.

The reduction in avoidance latencies is due to two factors. First, suppose that the increase in strength of the probability of crossing was about the same for each d/on/off state when one performed the avoidance response in that state. Over time, this would result in a decrease in avoidance latencies: As the probability of crossing in the early d/on/off states increased, the probability of an early latency would increase, even if the probability of crossing in later d/on/off states was equally high. This plays a role in the decrease in avoidance latencies, but it is not the whole story. Figure 12 shows the probability of crossing for each state throughout Simulation 2. The probability starts by being higher in the later d/on/off substates and only becomes high for earlier d/on/off substates later in training. This occurs because in temporal-difference models, value [$V(s)$] gradually propagates backward in time (Montague et al., 1996; Niv, Duff, & Dayan, 2005; Schultz et al., 1997). The values of states are updated by the prediction error (Equation 2). In the first trial, only the state that immediately precedes the aversive US gets a negative prediction error, when the system transitions from that state (with a value of 0) to a state that has a negative value (i.e., is aversive). Thus, in the first trial, only the last state before the US gets a negative value. In the second trial, the second state before the US also gets a negative prediction error, when the system transitions from that state (with a value of 0) to the last state before the US, which now has a negative value. Thus, in the second trial, some of the negative value propagates backward to the second state before the US. In addition, because learning is gradual, the last state before the US will still get a negative prediction error, so it will become even more negative. This process continues in subsequent trials, resulting in the gradual propagation of value backward. Before value propagates all the way back, it is more negative for late than for early substates of d/on/off. Thus, earlier in training, avoidances with longer latencies are more strongly reinforced than avoidances with shorter latencies, so avoidances will tend to have long latencies. As training progresses, value propagates back to the early substates of d/on/off, so early avoidances will also become strongly reinforced and the avoidance latencies decrease.

But why do avoidance latencies continue to decrease even as fear is markedly reduced? The explanation again

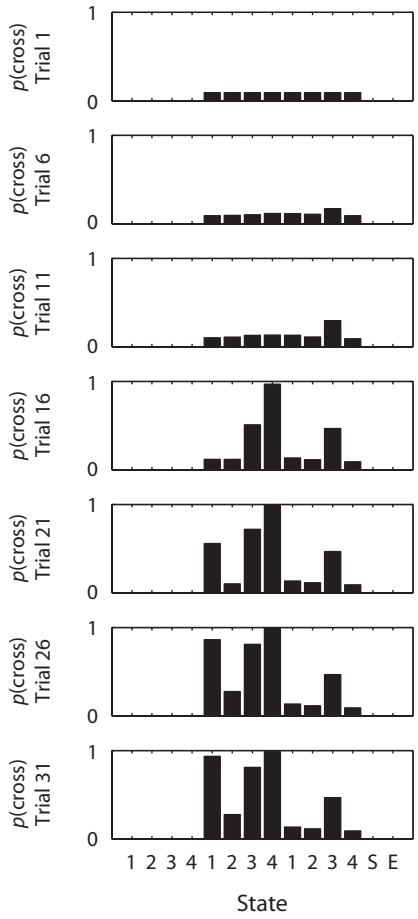

Figure 12. The response gradually transfers backward. These graphs show the probability of crossing for each state in Simulation 2, at 5-trial intervals. For simplicity, only the first 30 trials are shown. Each row shows these probabilities before the trial indicated on the vertical axis. During the first 10 trials (rows 2 and 3) the probability of crossing is greater for the d/on/on states than for the d/on/off states, so escapes are more prevalent than avoidances. (In this simulation, one specific d/on/on substate has an especially large probability of crossing. Different simulations would result in different d/on/on substates having large probabilities for crossing at this stage.) Starting in row 4, the probability of crossing in the d/on/off states has increased dramatically; in particular, the probability of crossing in the last d/on/off substate is nearly 1. Consistent with this, at this stage, the model already successfully avoids in every trial. However, in row 4 the probability of crossing is much lower for early than for late d/on/off substates; therefore, at this stage, avoidance responses still have long latencies. As training progresses, earlier substates of d/on/off also acquire a large probability of crossing; avoidance latencies therefore decrease.

hinges on the prediction error. As long as fear of a state has not fully extinguished, performing an avoidance in that state results in a positive prediction error, which reinforces the avoidance response. When this happens for the earlier d/on/off states, the avoidance latencies decrease. Referring back to Figure 10, the left column shows that the value for most states tends to extinguish during the extinction phase (and the same is true with extended acquisition). However, while the states have at least a somewhat negative value, the avoidance response in those states can continue to be strengthened. For example, after the end of acquisition (top row of Figure 10), the state d/on/off/1 has a negative value; until that value fully extinguishes, performing the avoidance response in that state results in a positive prediction error. This is why the probability of the response in that state increases markedly from Trial 15 to Trial 30 (Figure 10)—that is, after the end of the acquisition phase.

DISCUSSION

The Actor–Critic and the Need for Continued Reinforcement to Sustain Responding

We saw that the actor–critic explains why the avoidance response persists in the absence of continuing reinforcement. In animals, continuing reinforcement is not necessary to maintain a response in avoidance, but in other instrumental conditioning paradigms it is. The same thing happens in the actor–critic. As a simple example, consider an appetitive paradigm in which the animal learns to press a lever in response to a stimulus S to receive food. Before learning, the model is not expecting to receive food when it presses the lever, so when it does, there is a positive prediction error, which reinforces the leverpressing. With learning, S comes to have a positive value, because it is followed by leverpressing and the resultant food delivery. Now, suppose one stops giving food after the leverpressing. Stimulus S predicts a positive value, so the omission of reward produces a negative prediction error that weakens the response. The leverpressing response will therefore weaken or even extinguish when one removes the reinforcer, as is found empirically. The actor–critic therefore explains both why continuing reinforcement is necessary to sustain responding in appetitive conditioning and why it is not necessary in avoidance.

The Persistence of the Avoidance Response Versus the Persistence of Habits

It is important to contrast the account of the persistence of avoidance responses developed in this article with the idea of habits (Dickinson, 1985). Many experiments have shown that habitual responses become autonomous from their goals, persisting temporarily even when the animal is no longer interested in their outcome (Adams, 1982; Dickinson, 1985, 1994). In these experiments, in a first phase, animals are trained to respond for food. In a second phase, food is devalued by satiating the animal or by pairing the food with illness. Under certain training conditions (e.g., overtraining), the food devaluation does not lead to an immediate and pronounced reduction in responding, even though the animals are no longer interested in the food (e.g., Adams, 1982).

These results have been interpreted as suggesting that whereas goal-directed actions may be under the control of a model-based reinforcement learning system, habits may be under the control of a model-free reinforcement learning system (Daw, Niv, & Dayan, 2005, 2006). Model-based systems use a model of the environment

in which the relation of actions to their outcomes is explicitly represented. The model is often an MDP (such as the one in Figure 4), which includes knowledge of how actions affect the transitions between states and of the reinforcements associated with each state. Actions are selected in these systems by traversing the model to look at the consequences of the available actions. If the outcome of an action is no longer of interest (e.g., after food devaluation), these systems will therefore immediately stop performing that action. In contrast, in model-free systems such as the actor–critic (as in standard S–R associations), the strength of actions is stored independently from their outcomes. As Daw and colleagues put it, in model-free systems the action preferences are cached estimates (Daw et al., 2005; Daw, Niv, & Dayan, 2006). Even if the animal is no longer interested in the outcome of a previously desirable action, the strength of the action remains unchanged. It is only when the animal performs the action again that a negative prediction error occurs, decreasing the strength of the action. Furthermore, since learning in these systems is usually gradual, it takes a few trials for the strength of the action to be substantially decreased, which explains why food devaluation does not have an immediate effect on animals' responses when such responses have become habitual (Daw et al., 2005; Daw, Niv, & Dayan, 2006).

The cached values eventually catch up with the new contingencies, though, so the model would stop performing the behavior. Similarly, animals that are allowed to continue to respond for the devalued food soon stop responding (Adams, 1982). This is different from what happens in avoidance, in which the response tends not to extinguish even with many extinction trials (Solomon et al., 1953). Accordingly, in the model the persistence of avoidance responses and the persistence of habits are due to different mechanisms. Habits persist temporarily while the cached values are updated; avoidance responses persist lastingly because there is never a negative prediction error to weaken the response.

Relation to Cognitive Theories of Avoidance

An influential theory of avoidance that can explain the high resistance of avoidance responses to extinction is Seligman and Johnston's (1973) cognitive theory. Seligman and Johnston proposed that in the course of avoidance learning, animals learn two different action–outcome (A–O) expectancies for the situation in which the CS is on: If they do not perform the avoidance response, they will get shocked, and if they do perform the avoidance response, they will not get shocked. Since they prefer no shock to shock, this explains why they perform the avoidance response. Seligman and Johnston's explanation for the persistence of the avoidance response is that once the animal is performing the response consecutively on every trial, it never has an opportunity to disconfirm the expectation that it will get shocked if it does not perform the avoidance response.

It is not clear, however, that Seligman and Johnston's (1973) theory can explain all of the findings that this article addressed using the actor–critic—for example, the gradual reduction of avoidance latencies with additional training. From a "cognitive" perspective, there seems to be no reason for the avoidance latencies to become shorter than necessary with additional training. In fact, Seligman and Johnston's description of one of the expectations formed by the animal was that "The animal expects that if he responds within a given time (r_t, where t is the length of the CS–US interval . . .), no shock . . . will occur" (p. 91). But then, what is the driving force behind the gradual reduction in avoidance latencies? If anything, one would assume that with additional training the animal would estimate the length of the CS–US interval better, so it would be less compelled to perform the avoidance earlier just to be safe.

Using the terminology of reinforcement learning, the cognitive theory of Seligman and Johnston (1973) corresponds to a model-based approach. For such an approach to avoidance, see Smith et al. (2004) and Smith, Becker, and Kapur (2005).

Feedback From the Avoidance Response and Conditioned Inhibition of Fear

Sensory feedback that accompanies the performance of the avoidance response becomes a conditioned inhibitor of fear (i.e., a safety signal) that may positively reinforce the avoidance response (Dinsmoor, 2001; Dinsmoor & Sears, 1973; Morris, 1974, 1975; Weisman & Litner, 1972). Positive reinforcement by safety signals provides an alternative to two-factor theory's assumption that response reinforcement is due to CS termination. Alternatively, both mechanisms may play a role in response reinforcement (Cicala & Owen, 1976; B. A. Williams, 2001).

Safety signals also provide an alternative explanation for the persistence of avoidance responses. According to the Rescorla–Wagner model (Rescorla & Wagner, 1972), if the CS has a negative weight (i.e., elicits fear) and the safety signal has an equal but positive weight (i.e., inhibits fear), the prediction of the aversive US is 0. When the animal successfully avoids, no shock is presented, so both the CS and the safety signal would maintain their weights. Thus, the positive value of the safety signal and/or the termination of the negative value of the CS could continue to reinforce the response perpetually.

This account can also explain most of the other findings discussed above. For example, fear of the CS would never completely extinguish because it would be protected from extinction by the safety signal (Chorazyna, 1962; Rescorla, 1968). As another example, blocking the response would elicit fear because it would remove the safety signal, thereby uncovering the fear of the CS.

An important difficulty for this account, however, is to explain the decrease in avoidance latencies throughout consecutive avoidance responses. As was discussed above, fear of the CS decreases markedly with consecutive avoidance responses (even when it is measured by tests such as conditioned suppression, in which the avoidance response is not performed and the safety signal is therefore absent). Since the positive value of the safety

signal and the negative value of the CS should be symmetric, the value of the safety signal likely also decreases significantly. Reinforcement for the avoidance response, whether from termination of the CS or from the safety signal, therefore decreases markedly with consecutive avoidances. Both when a response is positively reinforced by appetitive stimuli (Crespi, 1942; Stebbins, 1962; Zeaman, 1949) and when it is reinforced by termination of an aversive stimulus (Strub, 1963; Woods, 1967), a decrease in reinforcement produces an increase in response latencies. In avoidance, however, one finds a decrease, not an increase, in response latencies throughout consecutive avoidances. It is unclear how this can be explained by this account.

Limitations of the Model

The present model has some limitations. In the model, all responses are treated equally. In animals, however, avoidance responses that are closer to species-specific defense reactions (SSDRs) are learned faster than arbitrary responses (Bolles, 1969, 1970; Grossen & Kelley, 1972). These findings initially led to the suggestion that reinforcement may not play any role in the establishment of avoidance responses (Bolles, 1970), but subsequent findings confirmed a role for reinforcement in avoidance (Crawford & Masterson, 1978, 1982).

In one-way avoidance, animals may quickly learn that one side is dangerous and the other is safe, and this may elicit a tendency to flee the dangerous side and approach the safe side (Bolles, 1978; Knapp, 1965; Zerbolio, 1968). Consistent with this hypothesis, if the dangerous and safe chambers are different (which facilitates learning about their values), one-way avoidance learning is faster than if the chambers are identical (Knapp, 1965). The tendency to flee the dangerous side and approach the safe side may also explain why learning is easier in one-way avoidance than in two-way or leverpressing avoidance (Bolles, 1972a). Nevertheless, the strengthening of the avoidance response throughout acquisition and even extinction (as evidenced by the reduction in latencies) suggests that reinforcement also plays a role.

The present article focused on one-way avoidance for simplicity. The goal, however, was to address general principles that apply across avoidance paradigms, so I did not model the noninstrumental tendency to flee the dangerous side and approach the safe side. Furthermore, pushing the envelope on a single explanatory construct (the standard actor–critic model) brings important theoretical insights. For example, the actor–critic can account for the finding of enhanced avoidance learning when the two chambers are different without postulating fleeing or approach processes. If the model used distributed representations for the state (Barto, 1995; Barto et al., 1983), there would be significant generalization of value across states; making the chambers more distinct would facilitate learning their values, which would produce faster learning. The model, like two-factor theory, can also explain why one-way avoidance is easier than two-way avoidance, which in turn is easier than leverpressing avoidance. In one-way avoidance, the animal moves to a context in which it has never been shocked, so there is a marked reduction in fear; in two-way avoidance, the animal moves to a context in which it was shocked recently, so there is less reduction in fear; and in leverpressing avoidance, the animal does not move out of the dangerous context, so there is even less reduction in fear (Bolles, 1978). Consistent with this explanation, if animals are allowed to move out of the dangerous context following leverpressing, learning of leverpressing avoidance is significantly faster (Masterson, 1970).

Reinforcement of the Avoidance Response by Prevention of the US

According to two-factor theory, the avoidance response is reinforced not because it prevents the US, but because it terminates the CS. Mowrer (1960) stated that "the avoidance of the shock is a sort of by-product" (p. 30) with no causal role in behavior. In the standard discriminated avoidance procedure, avoidance of the US and termination of the CS are confounded because the response has both effects. However, if the response prevents the US but does not terminate the CS (Bolles, Stokes, & Younger, 1966; Kamin, 1956), animals still learn it (albeit not as well as when it has both effects). This finding is usually taken to imply knowledge of the A–O contingency between the response and no shock.

The actor–critic does not represent A–O contingencies; for that, a model-based approach would be required. In fact, the brain may implement both model-based and model-free reinforcement learning (Daw et al., 2005; Daw, Niv, & Dayan, 2006), so both types of controller may play a role in avoidance. However, the finding that prevention of the US without termination of the CS is sufficient to support learning can also be given an interpretation on the basis of the actor–critic alone. States in reinforcement learning are not limited to representing external stimuli; they may also include, for example, memories (Sutton & Barto, 1998). Suppose that the states include memory of whether the avoidance response was executed. If the response prevents shock, the state that represents the fact that the response has been executed will eventually acquire a value of 0, because it is never followed by shock; the state that represents the fact that the response has not been executed, in contrast, will have a negative value, because it is followed by shock. Performing the response will therefore take the model from the negative (i.e., fear-eliciting) state that corresponds to the response not having been executed to the zero-valued state that corresponds to the response having been executed; this produces a positive prediction error that reinforces the response.

The relative contributions of CS termination and US prevention to avoidance learning may depend on the type of avoidance conditioning procedure (Bolles, 1970; Bolles & Grossen, 1969; Bolles et al., 1966). On the basis of two-factor theory and the model, one might expect that CS termination would play a smaller role in one- than in two-way avoidance, because in one-way avoidance the change from the dangerous to the safe chamber would be sufficient to produce substantial fear reduction. This has indeed been observed empirically (Bolles,

1969, 1970). By the same token, two-factor theory and the model would predict that CS termination should play a larger role in running in a wheel than in either one- or two-way avoidance, because when running in a wheel the animal does not change context at all, so fear reduction would be largely determined by the CS termination. That is not the case, though: CS termination does play a larger role in the running wheel than in one-way avoidance, but it appears to play a smaller role in the running wheel than in two-way avoidance (Bolles, 1970; Bolles et al., 1966).

SSDR theory attempts to explain these findings by claiming that the responses involved are different, with running in one-way avoidance and in the wheel being closer to an SSDR (Bolles, 1970). However, the response in two-way avoidance is also running, so appealing to differences in the responses themselves does not seem to explain much. One might argue that the SSDR is not running per se, but running to a safe place. But then why would running in the wheel be close to an SSDR, if it does not involve moving to a safe place? SSDR theory actually has the same difficulties as two-factor theory or the model in explaining this pattern of findings; additional research is needed to make sense of them.

Variability in the Resistance of Avoidance to Extinction

As was noted above, many studies have shown remarkable persistence of the avoidance response (Levis, 1966; Levis et al., 1970; Levis & Boyd, 1979; Logan, 1951; Malloy & Levis, 1988; McAllister et al., 1986; Seligman & Campbell, 1965; Solomon et al., 1953; Solomon & Wynne, 1953; Wahlsten & Cole, 1972; R. W. Williams & Levis, 1991). For example, Solomon et al. (1953) reported that with 200 extinction trials none of the dogs in their experiment extinguished the response. Three additional dogs run for 280, 310, and 490 extinction trials also failed to extinguish the response. With humans, R. W. Williams and Levis (1991) reported that nearly half of their subjects failed to extinguish after 500 trials, even though in many cases they had received a single pairing of the CS with a mild shock. Several studies, however, have reported gradual extinction of the avoidance response (see Mackintosh, 1974, for a review), even though there is often significant variability across animals (e.g., Sheffield & Temmer, 1950). Even in the R. W. Williams and Levis study, approximately half of the subjects *did* extinguish the response.

The model can explain the different susceptibilities to extinction across subjects if one makes two reasonable assumptions: (1) Different subjects have different values of the temperature parameter, τ, and (2) performing the avoidance response carries some cost (e.g., an energetic cost). Subjects with lower values of τ (i.e., those with a decreased tendency to explore) will tend to always avoid and therefore never discover that shock is no longer being delivered; subjects with higher values of τ are more likely to fail to perform the avoidance and therefore to discover that shock is no longer being delivered. Lower values of τ therefore produce more resistance to extinction. Consistent with this hypothesis, a strain of rats that performed the avoidance response on nearly every trial during acquisition showed more resistance to extinction than a strain that did not avoid as consistently (Servatius, Jiao, Beck, Pang, & Minor, 2008).

Why would the avoidance response extinguish in the model if τ is sufficiently high for the model not to avoid on every trial? Suppose that the avoidance has a cost c. Then, the states that precede it (those corresponding to the CS) will have a negative value ($-c$, if we ignore the discount factor γ for simplicity), even after fear of the CS is fully extinguished. Thus, when the model does not avoid and does not get shocked, it gets a positive prediction error c, because it goes from a state with value $-c$ to a state with value 0 (no shock). This has two consequences: (1) Not avoiding (i.e., performing whatever other actions the animal performed) is reinforced, and (2) the value of the CS becomes less negative ($-c + \alpha c$; see Equation 2). Then, when the avoidance response is performed again, it produces a negative prediction error $-\alpha c$, because the model was in a state in which it expected $-c + \alpha c$ and it gets $-c$. This negative prediction error weakens the avoidance response. Repetition of the weakening of the avoidance response and strengthening of alternative responses would eventually extinguish the avoidance response.

This discussion suggests that some of the variability across experiments may be due to different costs for the avoidance response. Other factors that may influence that variability include the intensity of shock (stronger shocks will produce larger prediction errors that stamp in the response more strongly) and the number of avoidance responses until the schedule is switched to an extinction schedule (fewer responses mean fewer opportunities for reinforcement and thus a weaker, more extinguishable response).

CONCLUSIONS

The actor–critic can be seen almost as a computational implementation of two-factor theory. The two components of the actor–critic (the actor and the critic) correspond closely to the two processes in two-factor theory (instrumental and classical conditioning, respectively). The actor–critic goes beyond two-factor theory, though. Unlike two-factor theory, the actor–critic does not predict that extinction of fear leads to extinction of the avoidance response. For this reason, the actor–critic can explain the persistence of avoidance responding, even after fear is greatly reduced or extinguished. In fact, the actor–critic even predicts the reduction in avoidance latencies during extinction trials. Importantly, the actor–critic allows us to understand why animals behave in the way that they do. For example, it explains why continuing reinforcement is not necessary to sustain responding in avoidance experiments but is necessary to sustain responding in appetitive paradigms. In short, the actor–critic explains the idiosyncrasies of the empirical findings in avoidance without postulating any mechanisms specific to avoidance.

AUTHOR NOTE

The author is now at Columbia University and the New York State Psychiatric Institute. This article is based on the author's doctoral dissertation in the Department of Psychology at Carnegie Mellon University. This work was supported in part by a Graduate Research Fellowship from the Calouste Gulbenkian Foundation. The author thanks James McClelland, John Anderson, Marlene Behrmann, and Ahmad Hariri for useful discussions about this work. Correspondence concerning this article should be addressed to T. V. Maia, Department of Psychiatry, Columbia University, 1051 Riverside Drive, Unit 74, New York, NY 10032 (e-mail: tmaia@columbia.edu).

REFERENCES

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

Baird, L. C. (1993). *Advantage updating* (Tech. Rep. No. WL-TR-93-1146). Dayton, OH: Wright-Patterson Air Force Base.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215-232). Cambridge, MA: MIT Press.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, & Cybernetics*, **13**, 835-846.

Beninger, R. J., Mason, S. T., Phillips, A. G., & Fibiger, H. C. (1980). The use of conditioned suppression to evaluate the nature of neuroleptic-induced avoidance deficits. *Journal of Pharmacology & Experimental Therapeutics*, **213**, 623-627.

Bolles, R. C. (1969). Avoidance and escape learning: Simultaneous acquisition of different responses. *Journal of Comparative & Physiological Psychology*, **68**, 355-358.

Bolles, R. C. (1970). Species-specific defense reactions and avoidance learning. *Psychological Review*, **77**, 32-48.

Bolles, R. C. (1972a). The avoidance learning problem. In G. H. Bower & K. W. Spence (Eds.), *The psychology of learning and motivation* (Vol. 6, pp. 97-145). New York: Academic Press.

Bolles, R. C. (1972b). Reinforcement, expectancy, and learning. *Psychological Review*, **79**, 394-409.

Bolles, R. C. (1978). The role of stimulus learning in defensive behavior. In S. H. Hulse, H. Fowler, & W. K. Honig (Eds.), *Cognitive processes in animal behavior* (pp. 89-108). Hillsdale, NJ: Erlbaum.

Bolles, R. C., & Grossen, N. E. (1969). Effects of an informational stimulus on the acquisition of avoidance behavior in rats. *Journal of Comparative & Physiological Psychology*, **68**, 90-99.

Bolles, R. C., Stokes, L. W., & Younger, M. S. (1966). Does CS termination reinforce avoidance behavior? *Journal of Comparative & Physiological Psychology*, **62**, 201-207.

Brady, J. V. (1965). Experimental studies of psychophysiological responses to stressful situations. In *Symposium on Medical Aspects of Stress in the Military Climate* (pp. 271-289). Washington, DC: Walter Reed Army Institute of Research.

Brady, J. V., & Harris, A. (1977). The experimental production of altered physiological states. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior* (pp. 595-618). Englewood Cliffs, NJ: Prentice Hall.

Bridle, J. S. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 211-217). San Mateo, CA: Morgan Kaufmann.

Chorazyna, H. (1962). Some properties of conditioned inhibition. *Acta Biologiae Experimentalis*, **22**, 5-13.

Cicala, G. A., & Owen, J. W. (1976). Warning signal termination and a feedback signal may not serve the same function. *Learning & Motivation*, **7**, 356-367.

Cook, M., Mineka, S., & Trumble, D. (1987). The role of response-produced and exteroceptive feedback in the attenuation of fear over the course of avoidance learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **13**, 239-249.

Coover, G. D., Ursin, H., & Levine, S. (1973). Plasma-corticosterone levels during active-avoidance learning in rats. *Journal of Comparative & Physiological Psychology*, **82**, 170-174.

Crawford, M., & Masterson, F. A. (1978). Components of the flight response can reinforce bar-press avoidance learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **4**, 144-151.

Crawford, M., & Masterson, F. A. (1982). Species-specific defense reactions and avoidance learning. An evaluative review. *Pavlovian Journal of Biological Science*, **17**, 204-214.

Crespi, L. P. (1942). Quantitative variation of incentive and performance in the white rat. *American Journal of Psychology*, **55**, 467-517.

Daw, N. D. (2003). *Reinforcement learning models of the dopamine system and their behavioral implications.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, **18**, 1637-1677.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, **8**, 1704-1711.

Daw, N. D., Niv, Y., & Dayan, P. (2006). Actions, policies, values, and the basal ganglia. In E. Bezard (Ed.), *Recent breakthroughs in basal ganglia research* (pp. 111-130). New York: Nova Science.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, **14**, 2567-2583.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, **36**, 285-298.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, **3**(Suppl.), 1218-1223.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B*, **308**, 67-78.

Dickinson, A. (1994). Instrumental conditioning. In N. J. Mackintosh (Ed.), *Animal learning and cognition* (pp. 45-79). San Diego: Academic Press.

Dinsmoor, J. A. (2001). Stimuli inevitably generated by behavior that avoids electric shock are inherently reinforcing. *Journal of the Experimental Analysis of Behavior*, **75**, 311-333.

Dinsmoor, J. A., & Sears, G. W. (1973). Control of avoidance by a response-produced stimulus. *Learning & Motivation*, **4**, 284-293.

Domjan, M. (2003). *The principles of learning and behavior* (5th ed.). Belmont, CA: Thomson/Wadsworth.

Estes, W. K., & Skinner, B. F. (1941). Some quantitative properties of anxiety. *Journal of Experimental Psychology*, **29**, 390-400.

Grossberg, S. (1972). A neural theory of punishment and avoidance. I: Qualitative theory. *Mathematical Biosciences*, **15**, 39-67.

Grossen, N. E., & Kelley, M. J. (1972). Species-specific behavior and acquisition of avoidance behavior in rats. *Journal of Comparative & Physiological Psychology*, **81**, 307-310.

Herrnstein, R. (1969). Method and theory in the study of avoidance. *Psychological Review*, **76**, 49-69.

Hodgson, R., & Rachman, S. (1974). II. Desynchrony in measures of fear. *Behaviour Research & Therapy*, **12**, 319-326.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249-270). Cambridge, MA: MIT Press.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. New York: Appleton-Century.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, **15**, 535-547.

Johnson, J. D., Li, W., Li, J., & Klopf, A. H. (2002). A computational model of learned avoidance behavior in a one-way avoidance experiment. *Adaptive Behavior*, **9**, 91-104.

Kamin, L. J. (1956). The effects of termination of the CS and avoidance of the US on avoidance learning. *Journal of Comparative & Physiological Psychology*, **49**, 420-424.

Kamin, L. J., Brimer, C. J., & Black, A. H. (1963). Conditioned suppression as a monitor of fear of the CS in the course of avoidance training. *Journal of Comparative & Physiological Psychology*, **56**, 497-501.

KLOPF, A. H., MORGAN, J. S., & WEAVER, S. E. (1993). A hierarchical network of control systems that learn: Modeling nervous system function during classical and instrumental conditioning. *Adaptive Behavior*, **1**, 263-319.

KNAPP, R. K. (1965). Acquisition and extinction of avoidance with similar and different shock and escape situations. *Journal of Comparative & Physiological Psychology*, **60**, 272-273.

LEVIS, D. J. (1966). Effects of serial CS presentation and other characteristics of the CS on the conditioned avoidance response. *Psychological Reports*, **18**, 755-766.

LEVIS, D. J., BOUSKA, S. A., ERON, J. B., & MCILHON, M. D. (1970). Serial CS presentation and one-way avoidance conditioning: A noticeable lack of delay in responding. *Psychonomic Science*, **20**, 147-149.

LEVIS, D. J., & BOYD, T. L. (1979). Symptom maintenance: An infrahuman analysis and extension of the conservation of anxiety principle. *Journal of Abnormal Psychology*, **88**, 107-120.

LEVIS, D. J., & BREWER, K. E. (2001). The neurotic paradox: Attempts by two-factor fear theory and alternative avoidance models to resolve the issues associated with sustained avoidance responding in extinction. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 561-597). Mahwah, NJ: Erlbaum.

LOGAN, F. A. (1951). A comparison of avoidance and nonavoidance eyelid conditioning. *Journal of Experimental Psychology*, **42**, 390-393.

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

MACKINTOSH, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, **82**, 276-298.

MAIA, T. V. (2007). *A reinforcement learning theory of avoidance.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

MAIA, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience*, **9**, 343-364.

MALLOY, P., & LEVIS, D. J. (1988). A laboratory demonstration of persistent human avoidance. *Behavior Therapy*, **19**, 229-241.

MASTERSON, F. A. (1970). Is termination of a warning signal an effective reward for the rat? *Journal of Comparative & Physiological Psychology*, **72**, 471-475.

MCALLISTER, W. R., & MCALLISTER, D. E. (1995). Two-factor fear theory: Implications for understanding anxiety-based clinical phenomena. In W. O'Donohue & L. Krasner (Eds.), *Theories of behavior therapy* (pp. 145-171). Washington, DC: American Psychological Association.

MCALLISTER, W. R., MCALLISTER, D. E., SCOLES, M. T., & HAMPTON, S. R. (1986). Persistence of fear-reducing behavior: Relevance for the conditioning theory of neurosis. *Journal of Abnormal Psychology*, **95**, 365-372.

MINEKA, S. (1979). The role of fear in theories of avoidance learning, flooding, and extinction. *Psychological Bulletin*, **86**, 985-1010.

MINEKA, S., & GINO, A. (1980). Dissociation between conditioned emotional response and extended avoidance performance. *Learning & Motivation*, **11**, 476-502.

MONTAGUE, P. R., DAYAN, P., & SEJNOWSKI, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, **16**, 1936-1947.

MORRIS, R. G. (1974). Pavlovian conditioned inhibition of fear during shuttlebox avoidance behavior. *Learning & Motivation*, **5**, 424-447.

MORRIS, R. G. (1975). Preconditioning of reinforcing properties to an exteroceptive feedback stimulus. *Learning & Motivation*, **6**, 289-298.

MOUTOUSSIS, M., BENTALL, R. P., WILLIAMS, J., & DAYAN, P. (2008). A temporal difference account of avoidance learning. *Network*, **19**, 137-160.

MOWRER, O. H. (1947). On the dual nature of learning—a reinterpretation of conditioning and problem solving. *Harvard Educational Review*, **17**, 102-148.

MOWRER, O. H. (1951). Two-factor learning theory: Summary and comment. *Psychological Review*, **58**, 350-354.

MOWRER, O. H. (1956). Two-factor learning theory reconsidered, with special reference to secondary reinforcement and the concept of habit. *Psychological Review*, **63**, 114-128.

MOWRER, O. H. (1960). *Learning theory and behavior*. New York: Wiley.

NEUENSCHWANDER, N., FABRIGOULE, C., & MACKINTOSH, N. J. (1987). Fear of the warning signal during overtraining of avoidance. *Quarterly Journal of Experimental Psychology*, **39B**, 23-33.

NIV, Y., DUFF, M. O., & DAYAN, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral & Brain Functions*, **1**, 6.

O'DOHERTY, J., DAYAN, P., SCHULTZ, J., DEICHMANN, R., FRISTON, K., & DOLAN, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**, 452-454.

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.

RACHMAN, S. (1976). The passing of the two-stage theory of fear and avoidance: Fresh possibilities. *Behaviour Research & Therapy*, **14**, 125-131.

RACHMAN, S., & HODGSON, R. (1974). I. Synchrony and desynchrony in fear and avoidance. *Behaviour Research & Therapy*, **12**, 311-318.

RESCORLA, R. A. (1968). Pavlovian conditioned fear in Sidman avoidance learning. *Journal of Comparative & Physiological Psychology*, **65**, 55-60.

RESCORLA, R. A., & WAGNER, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

RICCIO, D. C., & SILVESTRI, R. (1973). Extinction of avoidance behavior and the problem of residual fear. *Behaviour Research & Therapy*, **11**, 1-9.

SCHMAJUK, N. A., & ZANUTTO, B. S. (1997). Escape, avoidance, and imitation: A neural network approach. *Adaptive Behavior*, **6**, 63-129.

SCHULTZ, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, **80**, 1-27.

SCHULTZ, W., DAYAN, P., & MONTAGUE, P. R. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593-1599.

SELIGMAN, M. E. P., & CAMPBELL, B. A. (1965). Effect of intensity and duration of punishment on extinction of an avoidance response. *Journal of Comparative & Physiological Psychology*, **59**, 295-297.

SELIGMAN, M. E. P., & JOHNSTON, J. C. (1973). A cognitive theory of avoidance learning. In F. J. McGuigan & D. B. Lumsden (Eds.), *Contemporary approaches to conditioning and learning* (pp. 69-110). Washington, DC: Winston.

SERVATIUS, R. J., JIAO, X., BECK, K. D., PANG, K. C., & MINOR, T. R. (2008). Rapid avoidance acquisition in Wistar–Kyoto rats. *Behavioural Brain Research*, **192**, 191-197.

SHEFFIELD, F. D., & TEMMER, H. W. (1950). Relative resistance to extinction of escape training and avoidance training. *Journal of Experimental Psychology*, **40**, 287-298.

SMITH, A. J., BECKER, S., & KAPUR, S. (2005). A computational model of the functional role of the ventral-striatal D2 receptor in the expression of previously acquired behaviors. *Neural Computation*, **17**, 361-395.

SMITH, A. [J.], LI, M., BECKER, S., & KAPUR, S. (2004). A model of antipsychotic action in conditioned avoidance: A computational approach. *Neuropsychopharmacology*, **29**, 1040-1049.

SOLOMON, R. L., KAMIN, L. J., & WYNNE, L. C. (1953). Traumatic avoidance learning: The outcomes of several extinction procedures with dogs. *Journal of Abnormal Psychology*, **48**, 291-302.

SOLOMON, R. L., & WYNNE, L. C. (1953). Traumatic avoidance learning: Acquisition in normal dogs. *Psychological Monographs*, **67**(Whole No. 354).

SOLOMON, R. L., & WYNNE, L. C. (1954). Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review*, **61**, 353-385.

STARR, M. D., & MINEKA, S. (1977). Determinants of fear over the course of avoidance learning. *Learning & Motivation*, **8**, 332-350.

STEBBINS, W. C. (1962). Response latency as a function of amount of reinforcement. *Journal of the Experimental Analysis of Behavior*, **5**, 305-307.

STRUB, H. (1963). *Instrumental escape conditioning in a water alley: Shifts in magnitude of reinforcement under constant drive conditions.* Unpublished master's thesis, Hollins University, Roanoke, VA.

SURI, R. E., BARGAS, J., & ARBIB, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, **103**, 65-85.

SUTTON, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3**, 9-44.

SUTTON, R. S., & BARTO, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. R. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497-537). Cambridge, MA: MIT Press.

SUTTON, R. S., & BARTO, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

TAKAHASHI, Y., SCHOENBAUM, G., & NIV, Y. (2008). Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in Neuroscience*, **2**, 86-99.

THORNDIKE, E. L. (1911). *Animal intelligence: Experimental studies*. New Brunswick, NJ: Transaction.

WAHLSTEN, D. L., & COLE, M. (1972). Classical and avoidance training of leg flexion in the dog. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 379-408). New York: Appleton-Century-Crofts.

WEISMAN, R. G., & LITNER, J. S. (1972). The role of Pavlovian events in avoidance training. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning*. New York: Academic Press.

WILLIAMS, B. A. (2001). Two-factor theory has strong empirical evidence of validity. *Journal of the Experimental Analysis of Behavior*, **75**, 362-378.

WILLIAMS, R. W., & LEVIS, D. J. (1991). A demonstration of persistent human avoidance in extinction. *Bulletin of the Psychonomic Society*, **29**, 125-127.

WILLIAMS, Z. M., & ESKANDAR, E. N. (2006). Selective enhancement of associative learning by microstimulation of the anterior caudate. *Nature Neuroscience*, **9**, 562-568.

WOODS, P. J. (1967). Performance changes in escape conditioning following shifts in the magnitude of reinforcement. *Journal of Experimental Psychology*, **75**, 487-491.

ZEAMAN, D. (1949). Response latency as a function of the amount of reinforcement. *Journal of Experimental Psychology*, **39**, 466-483.

ZERBOLIO, D. J., JR. (1968). Escape and approach responses in avoidance learning. *Canadian Journal of Psychology*, **22**, 60-71.

NOTES

1. Note that fear in Figure 6 is a negative value because it is the prediction of an aversive outcome, and aversive outcomes are represented with negative scalars. Overall fear of the CS in Figure 8, in contrast, is the negative of the sum of the negative values for each state (see Equation 5), so it is positive.

2. The avoidance response becomes well established by Trial 12 in both simulations because during the first 15 trials they are equivalent. (Even though there is an element of randomness in action selection in the model, to make the simulations directly comparable I initialized them with the same random seed.)

(Manuscript received January 13, 2009;
revision accepted for publication August 7, 2009.)