Journal of Experimental Psychology: General
2006, Vol. 135, No. 2, 184–206

Copyright 2006 by the American Psychological Association
0096-3445/06/$12.00 DOI: 10.1037/0096-3445.135.2.184

From Recurrent Choice to Skill Learning: A Reinforcement-Learning Model

Wai-Tat Fu and John R. Anderson
Carnegie Mellon University

The authors propose a reinforcement-learning mechanism as a model for recurrent choice and extend it to account for skill learning. The model was inspired by recent research in neurophysiological studies of the basal ganglia and provides an integrated explanation of recurrent choice behavior and skill learning. The behavior includes effects of differential probabilities, magnitudes, variabilities, and delay of reinforcement. The model can also produce the violation of independence, preference reversals, and the goal gradient of reinforcement in maze learning. An experiment was conducted to study learning of action sequences in a multistep task. The fit of the model to the data demonstrated its ability to account for complex skill learning. The advantages of incorporating the mechanism into a larger cognitive architecture are discussed.

Keywords: reinforcement learning, recurrent choice, skill learning, credit assignment, cognitive architecture

One of the central issues in psychology concerns how people make choices from sets of alternatives. In some cases making choices involves a great deal of deliberation, even to the point of seeking professional advice and appointing subcommittees to assess different aspects of a choice. However, most of people's everyday behavior involves selections of particular actions out of all those that are physically possible, as they decide what to say next, whether to click on an option in searching the World Wide Web (Fu & Pirolli, in press; Pirolli & Fu, 2003), or whether to switch the channel on a TV set. The "selection" of actions does not imply a conscious or deliberative process. It simply refers to the fact that if the individual follows one particular course of action, there are other courses of action that he or she thereby forgoes. Such choices are not unique to humans, either—many creatures constantly make such choices in foraging in the wild or choosing among options in the laboratory. Our concern is with understanding these quick selections, typically made in less than one second, that determine the moment-to-moment course of people's lives. The ability to smoothly and accurately make such decisions is a major part of performing complex skills. Our interest arises out of the fact that the mechanism underlying such decisions plays a major role in cognitive architectures, and we cast our ideas in a production-system framework (Anderson et al., 2004; Meyer & Kieras, 1997a, 1997b; Newell, 1990). Part of what is new is that we incorporate recent ideas from research on reinforcement learning (e.g., Barron & Erev, 2003; Erev, Bereby-Meyer, & Roth, 1999; Gray, Sims, Fu, & Schoelles, in press; Holroyd & Coles, 2002; Houk, Adams, & Barto, 1995, Sutton & Barto, 1981), which is concerned with just these types of decisions.

At the outset, we would like to stress that there is a type of decision making to which our theory is not applicable. This is decision making that is accompanied by a time-consuming process characterized by "indecisiveness, vacillation, inconsistency, lengthy deliberation, and distress" (Busemeyer & Townsend, 1993, p. 432). Rather, we are concerned with the kind of quick, nondeliberative decision making that builds up through repeated exposures to simple situations. Although our theory is not applicable to the overall time-consuming process involved in the purchase of a house, it might describe little components of the purchase of a house, such as one's decision to smile when talking to a realtor or whether one looks at the dining room first or the living room. This is not to deny the importance of the large-scale, deliberative kind of decision making or to deny that there are some outcomes in common at the two timescales. However, this is not the concern of the current study; rather, our concern is with the outcome of a nondeliberative, statistical learning process over many exposures to the same or similar small-scale decisions. We discuss how this kind of learning is central to behavior from simple recurrent choice to complex skill learning shortly.

There have been numerous studies on this kind of nondeliberate choice under uncertainty in the last 50 years. Studies have been conducted on humans, pigeons, rats, and other creatures to understand how repeated rewards and punishments determine choice behavior in various situations (see Myers, 1976; Vulkan, 2000; and Williams, 1988, for reviews). In a typical experimental setting, an organism is asked to give repetitive responses such as button pressing or key pecking. Responses are intermittently rewarded or penalized according to a predefined probability. The recurrent choice behavior among the alternatives is found to be sensitive to the probabilities, the amount, and the variability of reinforcement (reward or penalty) as well as other schedule parameters such as

Wai-Tat Fu and John R. Anderson, Department of Psychology, Carnegie Mellon University.

This work was supported by Grant N00014-99-1-0097 from the Office of Naval Research. We thank Samuel McClure, James Houk, Nathaniel Daw, Peter Dayan, and Jennifer Ferris for their useful comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Wai-Tat Fu, who is now at the University of Illinois at Urbana–Champaign, 2414 Beckman Institute, 405 North Mathews Avenue, Urbana, IL 61801. E-mail: wfu@uiuc.edu

the time delay before reinforcement is given. One of the chief principles characterizing the allocation of choices has been the *matching law* (Herrnstein, 1961) and the related but different *probability matching*. Whereas probability matching refers to the general tendency to choose an alternative a proportion of the time equal to its probability of being correct, the matching law is more specific in terms of the mathematical relation between the proportion of choices and the proportion of the reinforcement. Specifically, the matching law is characterized as asserting that under steady-state conditions, the proportion of choices of an alternative is equal to the proportion of the total reinforcement allocated to that alternative.

Matching is often referred to as a molar measure, in the sense that it is largely independent of the moment-to-moment behavior of the organism and is often reached as a stable end state under particular conditions of reinforcement. The matching law says nothing about the individual processes constituting the molar measures. Recently, a good deal of effort has gone into developing molecular behavioral principles of the underlying process or processes to provide a moment-to-moment explanation for recurrent choice. On the one hand, psychologists have conducted experiments to identify a wide range of essential properties of recurrent choice in animals and humans, and specific models have been constructed to provide explanations of these properties. On the other hand, recent advances of neurophysiological studies of the basal ganglia have revealed important features of the reinforcement circuitry responsible for reward-related learning behavior. One of the purposes of this article is to build on previous work in these domains by constructing a unified, molecular mechanism that produces the diverse set of recurrent choice behavior. The other purpose of this article is to extend this recurrent choice mechanism to explain the acquisition of complex action sequences in skill learning.

Reinforcement learning is a process by which organisms learn from their interactions with the environment to achieve a goal or to maximize some form of reinforcement. The main characteristic of reinforcement learning is the use of a scalar reinforcement signal[1] that provides information about the magnitude of the “goodness” of an action in a particular context. Recently, it has been found that a reinforcement signal is provided by the firing patterns of dopaminergic neurons in response to sensory stimuli and the delivery of reward. Although much research has investigated reinforcement-learning mechanisms in neuroscience and machine learning, to our knowledge, no attempt has been made to test the capability of the mechanism to explain complex behavioral data. The primary goal of this article is to extend the scope of this neurally inspired reinforcement-learning mechanism to explain a breadth of recurrent choice and skill-learning phenomena. To achieve this goal, we identified a representative set of behavioral data that highlight the major phenomena in the literature. We then tested our version of the reinforcement-learning mechanism against these selected data sets. We believe that our results will help to bridge the gap between neurophysiological and psychological research on recurrent choice behavior. By putting the constraints from various areas together, we hope to produce an integrated theory that explains both recurrent choice and skill-learning behavior that can be implemented in a cognitive architecture such as ACT–R (Adaptive Control of Thought—Rational; Anderson & Lebiere, 1998).

One of the major strengths of the current model is its ability to learn in multistep recurrent choice situations, in which reinforcement is received after a sequence of choices have been made. From our review of the literature, we found that previous research on recurrent choice has almost exclusively used single-step choice tasks (one exception is the concurrent-chain schedule; e.g., Mazur, 2002) in which reinforcement is received after a choice is made in a single context. Because reinforcement is delayed in multistep choice tasks, a major difficulty is for the organism to determine the *critical* choices, among the sequence of choices made, that are responsible for the delayed reinforcement. Indeed, this problem, often called the *credit-assignment problem*, is central to the acquisition of complex action sequences in skill learning, in which one must make a sequence of action selections before feedback is received. Although numerous attempts have been made to tackle this problem in the machine learning literature, few attempts have been made to directly study how humans or animals learn to assign credit or blame to different actions. We therefore designed a general skill-learning task that studied how human subjects learn to acquire complex action sequences with delayed feedback. The results allowed us to test how well the reinforcement-learning mechanism can scale up to account for skill learning. To our knowledge, no existing psychological models have attempted to predict such sequential choice data.

In the animal learning literature, maze learning is a good example of a multistep choice problem. Maze learning has a long history in experimental psychology, and it has provided some of the strongest evidence for animals’ ability to adapt to the reward structures of complex environments. A number of studies have shown that rats are skillful at learning to navigate to the locations of food and other objects in complex mazes (e.g., O’Keefe & Nadel, 1978; Reid & Staddon, 1998; Spence, 1932; Tolman & Honzik, 1930). Although part of the learning depends on the development of “cognitive maps” (O’Keefe & Nadel, 1978), there is also evidence that rats learn to associate distinct cues in the environment to specific turns in the maze (Hull, 1934). This essentially involves the ability to apply the appropriate actions at different points in time. Part of the knowledge seems to result from the ability to associate cues in the environment to actions that will lead to some form of delayed reinforcement (Killeen, 1994; Machado, 1997; Reid & Staddon, 1998), an ability that is fundamental to the learning of most complex skills (see, e.g., Lewis & Anderson, 1985). A major contribution of this endeavor is to show that the reinforcement-learning mechanism that accounts for learning in simple recurrent choice can be extended to account for learning in complex skills.

This article has two major sections. In the first section, we develop the basis for our reinforcement-learning mechanism. We start with a brief review of the various properties of the reinforcement-processing circuitry in the cortical–basal–ganglionic loop, especially the relation between the dopaminergic signals in the basal ganglia and the temporal-difference error signals in reinforcement learning. This has served as the basis for a number of recent proposals (see, e.g., Holroyd & Coles, 2002; Yeung, Botvinick, & Cohen, 2004). Then, we develop the properties of our learning mechanisms on the basis of the ideas from this circuitry, from the machine learning literature, and from the

[1] A scalar signal has a single value and does not inform whether it is “good” or “bad”; that is, it contains the magnitude but not the valence information.

experimental psychology literature. Our proposal will be somewhat different from others in order for it to work naturally in a production-system framework.

The second section consists of various tests of our proposed mechanisms. First we test them against a diverse set of published data. Research on recurrent choice behavior has been studied for many years in a variety of procedures. To summarize, these procedures are designed to study the effects of one or more of the following four variables: the probabilities of receiving a reward, the magnitudes of the rewards, the variabilities of the rewards, and the delay of the rewards. We selected data sets that are representative in illustrating the effects of one or more of the above four variables. We were particularly interested in data that had been modeled by others and those that show not only the stable end states of choice allocations but also the learning trajectory that leads to these end states. These data sets allowed us to test whether the model could exhibit the main effects of each factor, as well as any interactions among them. We provide a detailed comparison between our model and the existing models in the General Discussion section.

Reinforcement Learning

The Basal Ganglia

The production system in ACT–R (Anderson & Lebiere, 1998) has been associated with the basal ganglia (Anderson, 2005; Anderson et al., 2004) on the basis of the circuitry of the basal ganglia and brain imaging data. Other researchers have identified the basal ganglia with reward-related learning. This hypothesis has been supported by two important architectural features: (a) the specialization of spiny neurons in the striatum for pattern recognition computations (Houk, 1995; Houk & Wise, 1995) and (b) the existence of relatively "private" feedback loops of connectivity from diverse cortical regions that converge onto those striatal spiny cells, via the pallidum and thalamus, and lead back to the frontal cortex (e.g., Alexander, Crutcher, & Delong, 1990; Amos, 2000; Kelly & Strick, 2004). The cortical–basal–ganglionic architecture creates a context-sensitive information-processing system that allows the striatum to instruct cortical areas as to which sensory inputs or patterns of motor outputs are behaviorally significant (see Figure 1). Unlike neurons that learn through a Hebbian-like mechanism, spiny neurons are found to receive specialized inputs that appear to contain training signals from dopamine (DA) neurons in the ventral tegmental and substantia nigra region (Schultz, Dayan, & Montague, 1997; Schultz et al., 1995). The availability of the training signals allows much more efficient learning, as dynamic information can be incrementally obtained from the environment.

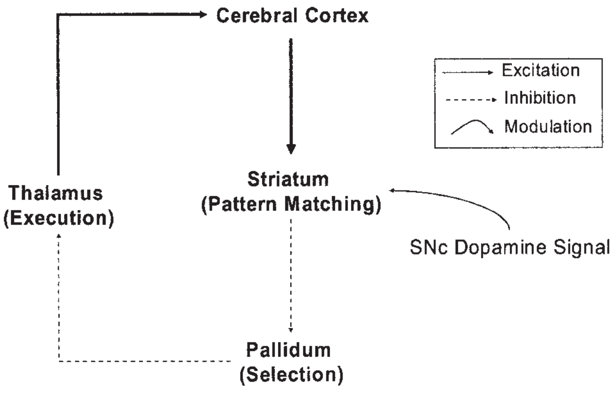


Figure 1. A simplified schematic diagram of the cortical–basal–ganglionic feedback loop. SNc = Substantia Nigra pars compacta.

Research has also found that when presented with reinforcement, the striatum appears to be capable of ordering its response in accordance with the valence (reward or punishment) and magnitude of the reinforcement (Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001; Delgado, Locke, Stenger, & Fiez, 2003; Yeung & Sanfey, 2004). For example, using a gambling paradigm, Delgado et al. (2003) found that the striatum differentiated between the valence (a "win" or "loss" event) for both large and small magnitudes of reward or punishment. In addition, a parametric ordering according to magnitude of reinforcement was observed in the left caudate nucleus: Activity of the striatum was the highest with large rewards, followed by small rewards, and the lowest with large penalties. Similarly, Yeung and Sanfey (2004), by measuring the event-related brain potential during a gambling task, found two distinct signals that were sensitive to the reward magnitude and the valence respectively, suggesting that reward magnitudes and valence were evaluated separately in the brain. These results seem to suggest that valence and magnitude of reinforcement may have differential effects on choice allocations.

Reinforcement information is believed to be carried by dopaminergic signals to the striatum. The role of the dopaminergic signals in the learning process was once widely believed to be signaling the occurrence of reward-related activities experienced by the organism. However, recent studies on the role of dopaminergic signals show that they do not simply report the occurrence of reinforcement. For example, Ljungberg, Apicella, and Schultz (1992) and Mirenowicz and Schultz (1994) showed that the activation of DA neurons depends entirely on the difference between the predicted and actual rewards. Once an unpredicted reward is perceived, response in DA neurons is transferred to the reward-predicting contextual patterns recognized by the striatum. Inversely, when a fully predicted reward fails to occur, DA neurons are depressed in their activity at exactly the time when the reward would have occurred (Pagnoni, Zink, Montague, & Berns, 2002; Schultz, Apicella, & Ljungberg, 1993). It therefore appears that outputs from DA neurons code for a deviation or error between the actual reward received and predictions or expectations of the reward. A simplified view[2] is that DA neurons seem to be feature detectors of the "goodness" of environmental events relative to the learned expectations about those events—that is, a positive signal is emitted when the reward is better than expected, no signal when the reward equals the expectation, and a negative signal when the reward is worse than expected. The output of the DA neurons is therefore often conceived of as an error signal fine-tuning the predictions of future reinforcement.

A number of researchers (Fu & Anderson, 2004; Holroyd & Coles, 2002; Houk et al., 1995; O'Reilly, 2003; Schultz et al.,

[2] The time course of the DA signals is more complicated (see Fiorillo, Tobler, & Schultz, 2003, 2005; Niv, Duff, & Dayan, 2005). Because our level of analysis is at the production level (i.e., approximately 50 ms), the complication is obviously outside the scope of this article.

1995, 1997) have associated the role of dopaminergic signals with the error signal in an algorithm called the *temporal difference* (TD) algorithm (Sutton & Barto, 1998). Sutton and Barto (1981) showed that although the TD algorithm has its roots in artificial intelligence, it can be easily related to animal learning theory. In fact, mathematically, it can be considered a generalization of the Rescorla–Wagner learning rule (Rescorla & Wagner, 1972) to the continuous time domain. Our learning mechanism is basically an elaboration of the TD algorithm that (a) makes it more suitable for modeling learning data and (b) incorporates it into a production-system framework.[3] In what follows we develop our application of the TD algorithm step by step.

Simple Integrator Model

The basic TD algorithm belongs to a class of learning models called simple integrator models (see, e.g., Bush & Mosteller, 1955). The simple integrator model is a popular model that has its form in discrete time as the following, also known as the *difference learning equation*:

$$V_i(n) = V_i(n-1) + a[R_i(n) - V_i(n-1)],$$

in which $V_i(n)$ represents the value or strength of some item i (e.g., a paired associate, stimulus–response bond, or production) after its nth occurrence and $R_i(n)$ represents the actual reinforcement (either a reward or a penalty) received on the nth occurrence. The parameter a ($0 < a < 1$) controls the rate of learning. One way to look at the model is that the prediction of the reinforcement, which usually reflects the strength of preference for a particular response, is updated according to the discrepancy between the actual reinforcement received and the last prediction of the reinforcement, that is, $R_i(n) - V_i(n-1)$. This discrepancy can be considered as an error signal, which drives the learning process that aims at minimizing the error signal; when $V_i(n-1) = R_i(n)$, the error is zero, and learning will stop. One appealing characteristic of this learning process is that the feedback takes the form of a scalar value without the valence information of "good" or "bad." As we will show later, this characteristic allows the model to learn from a continuous range of outcomes without the requirement of a "teaching signal" that informs whether the actions executed are good or not.

To understand the dynamics of the learning mechanism, consider the simple situation where a constant reward R is received in every time step. After t time steps, the expected reward can be shown to be $V_i(n) = R - (1-a)^t[R - V(0)]$, where $V(0)$ is the initial value of V. The integrator model therefore approaches the actual reinforcement value R with an exponential learning curve. The model, however, is quite limited as it depends only on the previous expectation and the current reinforcement. In cases where a sequence of actions is made before reinforcement is received, the effects of reinforcement need to be distributed across the sequence of actions. To solve this problem, we need a more general algorithm.

The TD Algorithm

The TD algorithm uses the same update equation as that in the simple integrator model. However, the definition of $R_i(n)$ is elaborated to include both the immediate reward, $r_i(n)$, and the value of the next item $i + 1$, $V_{i+1}(n-1)$:

$$R_i(n) = r_i(n) + g(t_i)V_{i+1}(n-1).$$

Note that this equation (the *discounted reward equation*) uses $n - 1$ to denote the value of the next item before it is updated. In this equation t_i is the time lag between item i and item $I + 1$; g is called the *discount function* that decreases monotonically with t_i. The farther away the future reinforcement is, the larger the discount will be and the less impact the value of the next item will have on the value of the current item. The expanded error term, $R_i(n) - V_i(n-1) = r_i(n) + g(t_i)V_{i+1}(n-1) - V_i(n-1)$, is called the *temporal difference* error.

One question concerns what the form of g should be. It can be shown that the sufficient condition for R to be bounded is that g is a monotonically decreasing function of t_i. In the original TD algorithm (Sutton & Barto, 1998), the exponential function is used as the discount function. Although the exponential function has nice computational properties, researchers have shown that it is often incapable of explaining results from empirical studies of delayed reinforcement (Ainslie & Herrnstein, 1981; Doya, Samejima, Katagiri, & Kawato, 2002; Mazur, 1984, 1985, 2001; Tadepalli & Ok, 1998). For example, Ainslie and Herrnstein (1981) compared the preference for a larger, later reward with a smaller, sooner reward and studied how this preference changed as a function of the delay to both rewards. They found that when the delay was small, subjects preferred the smaller, sooner reward, but as the delay increased, their preferences reversed in favor of the larger, later reward. It was found that the exponential function cannot predict the preference reversal. On the other hand, the hyperbolic function [$g(t) = 1/(1 + kt)$] predicts the preference reversal. Further support for the hyperbolic function was provided by Mazur (1984, 1985, 2001) on animal choice behavior and by Loewenstein and Prelec (1991) on human choice behavior. These studies have shown that the hyperbolic function provides good descriptions of choice behavior over extended periods of time. Our goal is to show how the discount function can be useful when incorporated in a more general learning mechanism that is capable of predicting local response patterns in a wide range of choice behavior.[4]

We can see that this learning rule updates prediction of future reinforcement by taking into account both the future reinforcement value and the time lag before the reinforcement is received. In

[3] However, it is not specific to our ACT–R theory, nor indeed has it yet been implemented in the ACT–R theory.

[4] It is worth noting that even if we make the discounting between successive items correspond to the hyperbolic function, the discounting across items may not be hyperbolic and thus may not predict preference reversal and the delay–amount tradeoffs. See Rachlin (2000, pp. 111–115) for partial support of this prediction. In general, if m items are repeatedly experienced and they are spaced at equal intervals t, then asymptotically,

$$V1 = \sum_{k=1}^{m} g(t)^{k-1} r(k),$$

where $V1$ is the value of the first production and $r(k)$ is the reward received at the kth production. In the special case where no reward is received until after the m items are presented, then $V1 = g(t)^{m-1}r(m) = [1/(1 + kt)]^{m-1}r(m)$. In this special case, the convergence properties associated with the standard exponential discounting will still be true with our hyperbolic discounting.

general, the reinforcement will have less impact on items that are farther away from it. This makes predictions about the order in which a sequence of choices will be learned. The choice close to the reinforcement will acquire value first, and then its value will propagate back to early choices. The idea of this kind of backward learning goes back at least to the Hullian notion of goal gradient (Hull, 1932). Hull stated that the goal event, often the reward, creates a "force field" that establishes a temporal gradient representing the distance to the goal. For example, in maze-learning experiments, the number of errors made by rats decreased when they were closer to the reward (e.g., Tolman & Honzik, 1930). Similarly, the theory of reinforcement of Killeen (1994) states that the effect of reinforcement has a "memory window," such that the reinforcement has the strongest influence on the response closest to the reinforcement and the influence on responses farther away from the reinforcement decays exponentially with time. A major contribution of our model is to propose a mechanistic account of how this kind of "backward spread of reinforcement" may influence recurrent choice behavior and, subsequently, complex skill-learning behavior.

Implementing a Reinforcement-Learning Mechanism in a Production-System Framework

Since their introduction by Newell (1973), production systems have had a successful history in psychology as theories of cognition (Anderson et al., 2004; Anderson & Lebiere, 1998; Klahr, Langley, & Neches, 1987). A production is basically a condition–action pair, with the condition side of a production rule specifying the state of the model (either internal cognitive state or external world state) and the action side specifying the action that is applicable at that particular state. Take, for example, the following two productions:

> *If* the goal is to choose between Button 1 and Button 2, *then* choose Button 1.
>
> *If* the goal is to choose between Button 1 and Button 2, *then* choose Button 2.

The "if" sides of the productions specify the state in which the model must choose between Button 1 and Button 2, and the "then" sides of the productions specify the action of choosing either one of the buttons. In ACT–R, each production has a utility value, which influences the likelihood of executing the production when it matches the current state of the model. The execution of a production will change the state of the model or the external world, which will lead to the next set of applicable productions. The next production is then selected and executed, and so on. Production systems have been applied to explain a broad range of cognitive phenomena, especially those that emerge from combinatorially complex tasks, such as studies of problem solving or human–computer interaction. As one can easily imagine, the behavior of a model depends critically on the process of selecting the next productions. In this section, we show how we implement the reinforcement-learning mechanism so that the model learns how to select productions to maximize its long-term reward. In subsequent sections we will show how we can use the same mechanism to account for behavior that spans from simple recurrent choice to complex skill learning.

A number of production systems have incorporated reinforcement-like learning. More or less standard reinforcement learning has been used for operator selection in the Soar architecture (Nason & Laird, 2004). ACT–R (Anderson et al., 2004) has a reinforcement-like learning mechanism for selecting among productions. However, none of the prior theories have the system we propose here, nor have these systems addressed the range of data of concern here (although Lovett, 1998, did address one of these phenomena within the ACT–R framework). We assume a general framework like that in ACT–R, where the reinforcement-learning mechanism applies to production rules and enables the selection among competing rules. The focus on learning the value of rules is part of a tradition in psychology that extends from stimulus–response (S-R) learning (and production rules can be seen as the modern embodiment of S-R bonds) to connectionist learning (where synapses are the equivalent of S-R bonds). Although many reinforcement-learning algorithms try to learn the value of states, our proposal is similar to the learning of state-operator transitions in the Q-learning algorithm (Watkins, 1989). In Q-learning, an agent tries to learn the values of actions in particular states—that is, it learns the value of state–action pairs. The learning of the values of productions is also consistent with the notion that reinforcement affects the environment–behavior relation, not just a response (Skinner, 1938).

As in the above example, in many choice situations of interest more than one production can be applied. Therefore, besides calculating the values of each production, we also need to define a policy to select among the productions. One of the most common policies in reinforcement learning is called the "soft-max" function (see, e.g., Sutton & Barto, 1981), which, coincidentally, is also a part of the ACT–R theory.[5] We adopted the ACT–R formulation of this soft-max function, in which the production with the highest value is selected, but these values are noisy. The noise is added from a normal-like logistic distribution. Thus, the production that will be selected can randomly vary from trial to trial. There is an approximate analytic equation (for details, see Anderson & Lebiere, 1998, chap. 3, Appendix A) that gives the probabilities for selecting any production in a conflict set of competing rules. If V_i is the evaluation of alternative i, the probability of choosing that ith production among n applicable productions with evaluations V_j will be

$$\frac{\exp(V_i/t)}{\sum_{j=1}^{n}\exp(V_j/t)},$$

where the summation is over the n alternatives. The parameter t in the above distribution is related to the standard deviation, σ, of the noise by the formula $t = \sqrt{6}\sigma/\pi$. This equation (the *conflict-resolution equation*) is the same as the Boltzmann equation used in Boltzmann machines (Ackley, Hinton, & Sejnowsky, 1985; Hinton & Sejnowsky, 1986). In this context t is called the temperature. This conflict-resolution equation has the property that as t approaches 0, the probability of choosing the production with the highest value approaches 1 (i.e., becomes deterministic). As t

[5] Although ACT–R has a very different assumption behind the use of the soft-max function, the mathematical form is basically the same.

increases, the probability of choosing the production with the highest value decreases. The equation therefore allows a balance between exploration and exploitation by varying the value of *t*—that is, it allows the model to explore other alternatives even when those alternatives may have low utility values. The smaller *t* is (and the less noise), the stronger will be the tendency to select the maximum item. Note that the conflict-resolution equation represents the expected long-term behavior of the selection process. The actual predictions depend on Monte Carlo simulations.

We illustrate our algorithm in a simple model with four productions that fire in cycles. During each cycle, the production fires in this sequence: P-Step1, P-Step2, P-Step3, and P-Step4. After P-Step4 is fired, a reward of 1 is received, and another cycle begins by firing P-Step1. Figure 2 shows how the expected values change with the reinforcement-learning mechanism. We used the parameters in this model that were used throughout the article—learning rate *a* was .05 and delayed-reward parameter *k* was .25. In Figure 2 we assume a 1-s delay between each production, which means that each production discounts the value of the next by 1/(1 + .25 * 1) = .8. After the 150 trials in Figure 2, the productions are reaching their steady-state values, which can be expressed as

$$V_1 = .8V_2 = 0.512$$

$$V_1 = .8V_2 = 0.512$$

$$V_2 = .8V_3 = 0.640$$

$$V_3 = .8V_4 = 0.800$$

$$V_4 = 1.0 + .8V_1 = 1.00$$

$$V_2 = .8V_3 = 0.640$$

$$V_3 = .8V_4 = 0.800$$

$$V_4 = 1.0 + .8V_1 = 1.00.$$

Consistent with the goal-gradient hypothesis, Figure 2 shows how the reinforcement signal propagates back in time and diminishes as it goes farther away from the actual reward. In later sections, we provide further examples that show how the discounting of delayed rewards explains skill-learning behavior.

Testing the Mechanism Against Empirical Data

We tested the predictions of the reinforcement-learning mechanism against several sets of data. To highlight the properties of the mechanism, we present the results in three parts. In Part 1, we chose data sets from simple recurrent choice situations where two or more alternatives are chosen repeatedly with different probabilities, magnitudes, and variabilities of reinforcement. We aimed at testing whether the combination of the basic TD algorithm (with properties of a simple integrator model) and the choice rule (the soft-max function) is capable of exhibiting the same sensitivities to the three major manipulations of reinforcement as human subjects. In Parts 2 and 3, we tested aspects of the model that are specific to our version of the TD algorithm. In Part 2 we tested the assumption of the hyperbolic discount function and how it predicted the various effects of the delay of reinforcement. In Part 3 we tested the process by which the TD algorithm propagates credits back to earlier productions and how it can account for skill-learning behavior. Throughout this effort we held all of the critical parameters in the theory constant:

$$\text{Delayed reward parameter } k = .25$$

$$\text{Learning rate } a = .05$$

$$\text{Noise parameter } t = 1.00$$

The setting of the noise parameter basically sets the scale for the values (and so really is not an estimated parameter). The only parameters we estimated to fit particular data sets were the reward parameters (*r*) associated with various outcomes. These parameters correspond to what the experimenter manipulated and so would be expected to vary from experiment to experiment and condition to condition. The initial expected values of all the productions were set to 0 at the beginning of the experiments and changed as a function of experience. In other words, we assumed equal preference for all choices a priori and focused on tasks that show how preferences change owing solely to the accumulation of reward experiences.

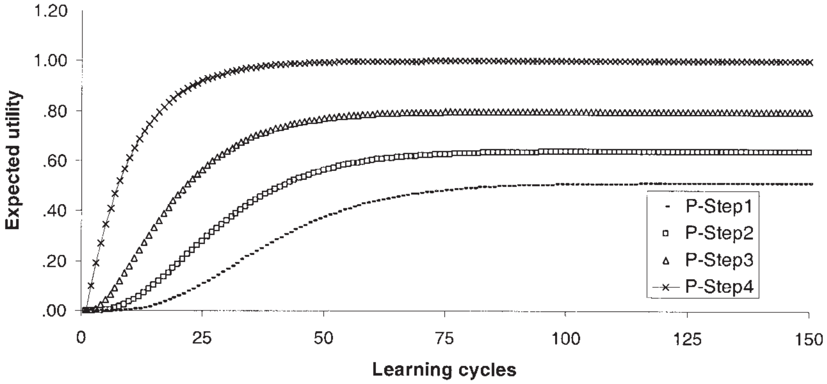


Figure 2. The expected values of the four productions that fired in cycles and received a reward of 1 after P-Step4 was fired. The time for each production to fire was 1 s. Learning rate (*a*) = .05; delayed-reward parameter (*k*) = .25.

Part 1: Learning From Differential Probabilities, Magnitudes, and Variabilities of Reinforcement

In the first part, we focus on three factors in recurrent choice: the probabilities, magnitudes, and variabilities of reinforcement. These three factors cover most of the literature on recurrent choice behavior (the other factor is the delay of reinforcement, which we discuss in Part 2). First, we selected one of the most comprehensive data sets by Friedman et al. (1964) on how people learn to guess which of two alternatives occurs—a paradigm called *probability learning*. Second, we selected the data set by Myers and Suydam (1964) that manipulated the magnitudes of reward orthogonally with the event probabilities. Many published data sets have shown the effects. We chose these two data sets because they provided the most comprehensive information on the experiments and the data, especially the fact that they presented both learning and asymptotic performance across a large number of trials. In addition, these data sets were modeled by others: The data set by Friedman et al. was modeled by Lovett (1998), and that by Myers and Suydam was modeled by Busemeyer and Townsend (1993) and Gonzalez-Vallejo (2002). Testing our model against these two data sets therefore allowed a direct comparison among models. For example, we will show that although Lovett's model can account for the data set by Friedman et al., it has trouble fitting the data set by Myers and Suydam.

We then tested our model in more complex situations where the payoffs were not fixed. We selected the data set by Busemeyer and Myung (1992) that showed how people learn to choose in situations that meet with a continuous range of outcomes with different probabilities. This data set is particularly problematic for models that require a "training signal" that informs the model to change its preferences for different options. We showed that our model describes the interesting patterns of behavior, and we compared our model with the adaptive network model of Busemeyer and Myung. Last, we selected the data set by Myers, Suydam, and Gambino (1965) that showed the interaction effects of differential payoffs and event probabilities and how they influenced risk-taking behavior. This data set is from a classic study, and many consider it a benchmark test for existing models (e.g., Barron & Erev, 2003; Busemeyer & Townsend, 1993; Gonzalez-Vallejo, 2002). In summary, the four data sets we selected in Part 1 showed how our model accounts for the effects of the major manipulations of reinforcement in the recurrent choice literature.

Probability Learning

A fairly direct form of choice under uncertainty is the study of how people select between multiple alternatives with uncertain outcomes and rewards. The simplest situation is the probability-learning experiment, in which a participant guesses which of the alternatives occurs and then receives feedback on his or her guesses. We start with the consideration of this paradigm, as the changes in choice proportions directly reflect how preferences change with reward experiences. Thus, it served as one of the bases for the development of the current reinforcement-learning mechanism.

In Friedman et al. (1964), participants completed more than 1,000 choice trials over the course of 3 days. For each trial, a signal light was illuminated, participants pressed one of the two buttons presented, and then one of the two outcome lights was illuminated. Task instructions encouraged participants to try to guess the correct outcome for each trial. The two buttons' success probabilities varied during each 48-trial block in the experiment. Specifically, for the odd-numbered blocks 1–17, the probabilities of success of the buttons (p and $1 - p$) were .50. For the even-numbered blocks 2–16, p took on the values from .10 to .90, in a random order. Therefore in the first block, the two buttons were equally likely to be correct. Starting from the second block and in each of the subsequent even-numbered blocks, one of the buttons was more likely to be correct. There were 48 trials in each block, and Friedman et al. provided the mean choice proportions of each button in every 12-trial subblock. This allowed us to match the learning of the event probabilities across the 48 trials in each block (i.e., the four 12-trial subblocks in each block). We focus on the analysis of the even-numbered blocks, as they show how people adapted to the different outcome probabilities with experience.

Figure 3a shows the predicted proportion of choices in the experiment by Friedman et al. (1964). Participants in general exhibited "matching" behavior—that is, they came to choose a button approximately in proportion to its probability of being correct. Across the four 12-trial subblocks in each of the even-numbered blocks, participants chose the correct button in roughly 50% of the trials in the first subblock and approached the corresponding p values by the final subblock. This is called *probability matching*.

Figure 3b shows the proportions predicted by the model, which had four productions: *Prepare*, *Button 1*, *Button 2*, and *Finish*. We use variations on this same model throughout this section. Before a trial starts, *Prepare* fires, waiting for the trial to begin. Then one of the productions, *Button 1* or *Button 2*, fires. If the button chosen is a correct button, a reward of value r will update the value of the production; otherwise, a reward of $-r$ will be received. The *Finish* production then fires, which leads to the start of the next trial. We assume that each of these productions takes about 1 s of processing. Given the 1-s delay and the value of $k = .25$, the discount value, $1/(1 + kt)$, is .8 between pairs of productions.

The expected values of the four productions were updated according to the difference learning equation in the reinforcement-learning mechanism. There was an r or $-r$ amount of reward credited to the productions that chose the buttons. To fit the data, we set r to 1.4. The exact sequence of outcomes as reported in Friedman et al. (1964) was presented to the model. We obtained a fit of $R^2 = .912$. The standard error of the model's estimate (a measure of deviation between prediction and observation) is .056. As one measure of the reliability of the data, we looked at the degree to which the probability of choice in the p condition could be predicted by 1 minus the probability of choice in the $1 - p$ condition. They should be the same because the two alternatives are indistinguishable. The average deviation in these numbers is .058. Therefore, it seems the model does correspond to the data within the limits of the data's own precision. The fit is also similar to that obtained by Lovett (1998). However, because the model by Lovett combined event probabilities and amount of reward into a single parameter, it has trouble fitting the data set by Myers and Suydam (1964), which we discuss shortly.

Figure 3c provides an illustration of the changes in the values that are behind these predictions. There we have plotted the change in the values of the four productions over the course of the 48 trials. It is assumed that each production has an initial value of 0, and we plot the average values of the productions after each

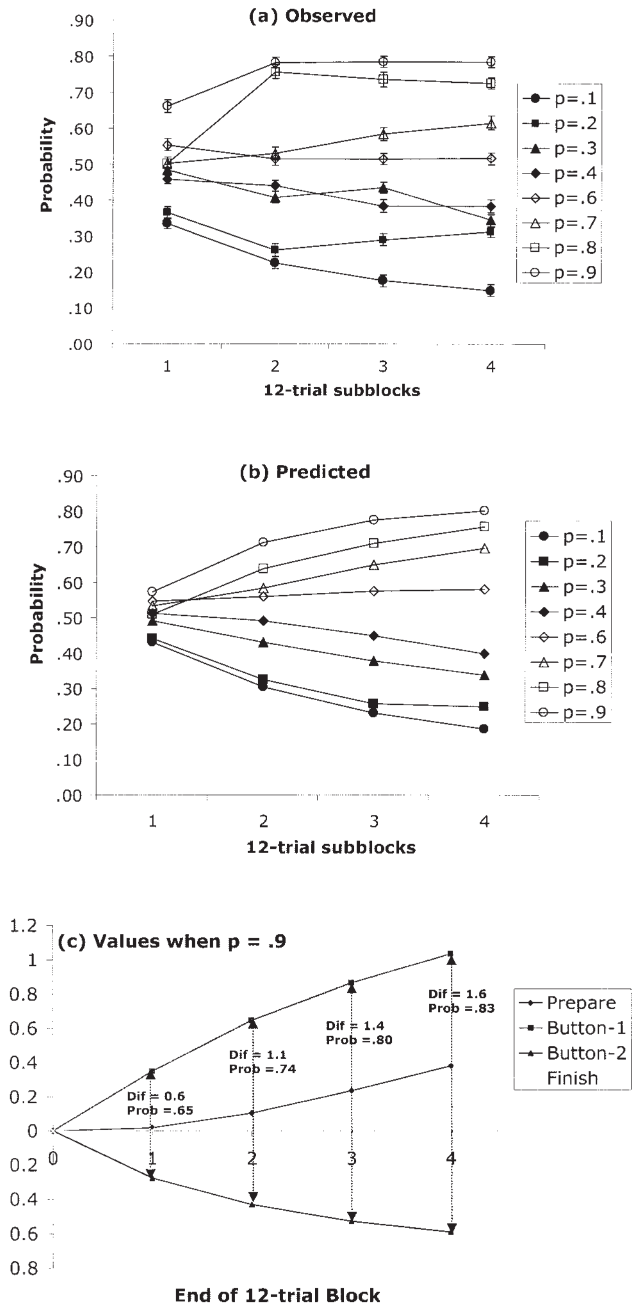


Figure 3. Observed (a) and predicted (b) choice proportions in Friedman et al. (1964) across four 12-trial blocks of different probability of success (p) for one of the buttons. Part (c) shows the average change in values across the course of the experiment in the condition when $p = .90$. Dif = difference; prob = probability.

subblock of 12 trials. It can be seen that the high-probability-of-reward button (Button 1) grows rapidly whereas the low-probability-of-reward button (Button 2) decreases fairly rapidly. The figure indicates the difference in values between the two button productions and how this difference maps onto probability of choice according to the conflict-resolution equation. As the model increasingly tends to choose the high reward button (Button 1), it passes the positive reinforcement signal back to the sequence of productions that preceded Button 1. As a result, the value of the Prepare production tends to increase. Similarly, the Finish production accrues .8 of the value of the Prepare production that follows it.

Contingent Payoffs

In the previous two experiments, the two outcomes were basically symmetrical in the sense that whichever option was predicted, if it occurred the same reward was given. However, one can also make payoffs contingent on which action is predicted. One such example is the study by Myers and Suydam (1964), which was also chosen as a benchmark test for the models by Busemeyer and Townsend (1993) and Gonzalez-Vallejo (2002). Myers and Suydam's experiment provides further evidence that a successful choice model must be sensitive to the probabilities and magnitudes of reward separately. Lovett's (1998) model would have difficulty with this experiment because it used a single parameter to represent the reward obtained from choosing the outcomes; in contrast, Myers and Suydam's design made the reward depend on the choice. In Myers and Suydam's experiment, if participants chose the first alternative (which was always the more probable) and were correct, they received a reward of G points; if they were wrong, they lost L points, and the values of G and L depended on condition. The second alternative always paid 1 point if correct and −1 point if wrong. The values of G and L (expressed as points) are shown in Table 1. All participants started with a stake of 100 points, redeemable at .25 cent per point. The study was a 2 (probability of Alternative 1 being correct = .60, .80) × 2 (G = 1, 4) × 2 (L = 1, 4) between-subjects design. The expected payoffs for alternatives in each condition are also shown in Table 1. For example, the expected payoff of Alternative 1 in the $p = .60$, G = 1, L = 1 condition can be calculated as $(.60)1 + (1 - .60)(-1) = .20$, and that of Alternative 2 will be $(.60)(-1) + (1 - .60)(1) = -.20$.

If preference strengths are sensitive to the expected payoffs, then preferences should be in the order of the differences between the two options, hence the following order: (G = 4, L = 1) > (G = 4, L = 4) > (G = 1, L = 1) > (G = 1, L = 4). It is interesting to note that except when $p(1) = .60$, G = 1, and L = 4, Alternative 2 should be chosen over Alternative 1. Figure 4 presents two displays of the results. Figure 4a shows the average proportion of choice of Alternative 1 over all 300 trials for the eight conditions of the experiment. The results were in general consistent with the predictions based on expected payoffs: Partic-

Table 1
Expected Payoffs (EP) for Each Condition in Myers and Suydam's (1964) Experiment

		P(1) = .60			*P*(1) = .80		
G	L	EP(1)	EP(2)	Dif	EP(1)	EP(2)	Dif
1	1	0.2	−0.2	0.4	0.6	−0.6	1.2
1	4	−1.0	−0.2	−0.8	0	−0.6	0.6
4	1	2.0	−0.2	2.2	3	−0.6	3.6
4	4	0.8	−0.2	1.0	2.4	−0.6	3.0

Note. G and L reflect points gained and lost, respectively. Dif = difference.

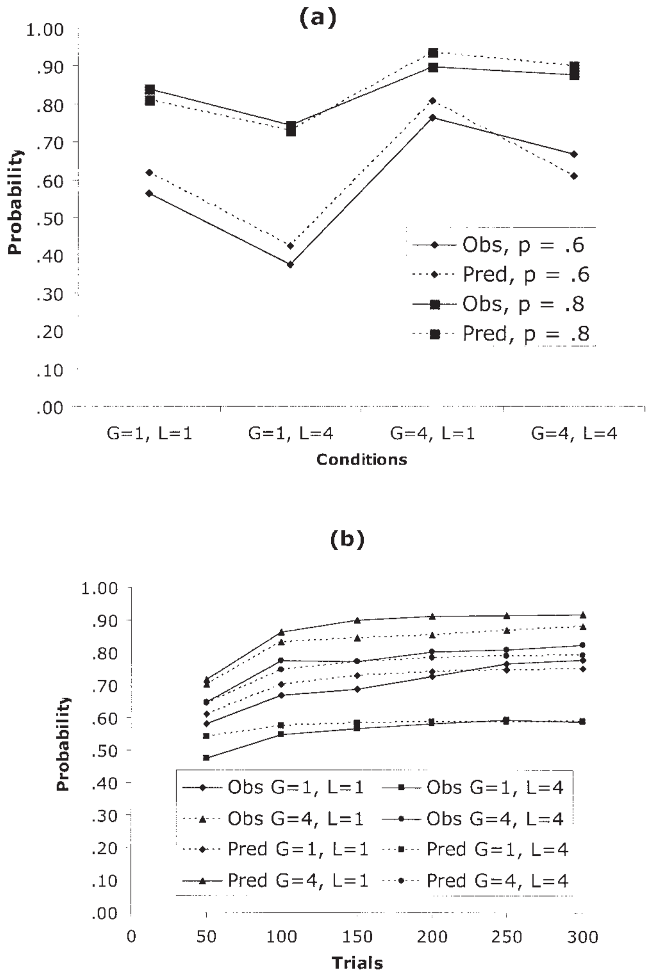


Figure 4. Observed (obs) and predicted (pred) choice proportions from the experiment by Myers and Suydam (1964) in the eight conditions (a) and learning rates (b). G and L indicate points gained and lost, respectively.

ipants chose Alternative 1 more often in all but the condition in which $p(1) = .60$, G = 1, and L = 4, and the magnitude of choice proportions was in the predicted order: (G = 4, L = 1) > (G = 4, L = 4) > (G = 1, L = 1) > (G = 1, L = 4). Myers and Suydam (1964) found a significant G × L interaction. Specifically, increases in L had a larger effect when G = 1 than when G = 4. Figure 4b shows averages over the .60 and .80 conditions and displays the average learning curves, in blocks of 50 trials, for each combination of G and L. These four learning curves are rather similar. Except for the G = 1, L = 1 curve, participants show about 2/3 of the rise from Block 1 to Block 2 that they show over the entire course of the experiment. In general, participants appear to have reached close to asymptote by the end of the experiment.

Figure 4 also displays the fit of our model to these data. We allowed ourselves to estimate independent rewards for each outcome: +4, +1, −1, and −4. Our estimates of these values were $R(+4) = 2.9$; $R(+1) = 1.2$; $R(-1) = -1.8$; and $R(-4) = -4.2$. As can be seen, the losses are of greater absolute magnitude than the positive rewards, corresponding to the frequent finding that people are loss averse.[6] All other parameters were the same as in previous models. We obtained a fit of $R^2 = .932$ (to all 48 observations yielded by crossing the 8 conditions with the 6 points on the learning curve). The standard error of the model's estimate was .048. Note in particular that the model captures the interaction between the levels of L and G noted by Myers and Suydam (1964) and the observed rate of learning. Although the estimates of parameters for magnitude of rewards will substantially determine the asymptotic levels of choice in the model, the predictions about the learning rate really fall out from the fixed parameters of the model. We should point out that our model is the first to produce these learning curves, although the average choice proportions in Figure 4a have been modeled by others (Busemeyer & Townsend, 1993; Gonzalez-Vallejo, 2002).

Learning From Variable Rewards

In the preceding experiments a particular choice was either right or wrong. However, in many situations a choice can meet with a continuous range of outcomes. Reinforcement learning is particularly suited for this type of learning because it does not require an explicit teacher who offers the correct answer. Instead, reinforcement is simply compared with existing expectation, and if it is higher than expectation, it is considered a positive reinforcement; otherwise, it is considered a penalty. Busemeyer and Myung (1992) conducted such an experiment in which participants were told to select one of the three treatment strategies for patients suffering from a common set of symptom patterns. Feedback on the effectiveness produced by the treatment was given after each selection. For the sake of convenience, the treatment with the highest expected effectiveness is called Treatment 3, the next most effective is called Treatment 2, and the least effective is called Treatment 1. Figure 5 illustrates the distribution of payoffs of the treatments. The effectiveness produced by each treatment was normally distributed with equal standard deviation, but the mean payoffs are different (as explained below). Participants had to evaluate each treatment on the basis of trial-by-trial feedback. Participants were told to maximize the sum of the treatment effects over training, and they were paid 4 cents per point. The means of the normal distributions are equally spaced apart for Treatments 1, 2, and 3. The two independent variables were mean difference (*d*) (i.e., the separation of the distributions in Figure 5) and standard deviation (which affects the amount of overlap in Figure 5). The exact values of *d* and the standard deviations are shown in Table 2. Each participant was given nine blocks (50 trials per block) of training in each condition. The model received the same amount of training as the participants.

From Table 2 we can see that as *d* increased, the observed choice proportions of the optimal treatment increased. As the standard deviation increased, the observed choice proportions of the best treatment decreased. These two main effects were significant, and the interaction between them was not. These results were quite striking, as they showed how participants estimated the payoff distributions on the basis of repeated sampling of the

[6] Asymptotically, the only thing that matters for prediction is the difference between the utilities and not their actual values. However, as the learned utilities all start at zero, the predicted learning curves depend somewhat on the actual values. Nonetheless, our predictions would change only modestly if we were to adjust these four utilities by adding a constant so that their average is zero.

effectiveness of each treatment without any information on which treatment was the optimal treatment.

To model the data, we used the same model as we did for the Friedman et al. (1964) data, except that three productions were used to choose each of the three treatments. Similar to all previous models, the initial expected value of each production was set to 0. For each trial, the rewards obtained by the model were simulated by drawing a sample from the normal distribution that represents the effectiveness of the treatment chosen by the model. One of the problems we faced was how to scale these objective rewards onto subjective rewards for the model. We could not take the approach of the previous model of simply estimating a different subjective reward value for each objective reward, because there are an infinite number of such rewards. Although we could have taken the approach of trying to estimate a more complex function, we simply used the following equation to map the objective reward onto the subjective reward value:

$$\text{Subjective reward} = \text{scale}*\text{objective reward}.$$

Curiously, we estimated the multiplicative scale factor to be exactly 1.00 to best fit the data. We obtained a fit of $R^2 = .782$. This is smaller than the previous fits; however, the correlation between the numbers and what would be predicted from the main effects alone is only .804, and the interaction term is not significant. This suggests that we may be fitting the data to the degree of its reliability. The standard error of the model's estimate was .039. The new learning mechanism built up the distributions of effectiveness of the treatments from trial-by-trial feedback and exhibited similar sensitivity to the differences of the variabilities and means of the distributions as participants. The average difference between standard deviations of 6.0 and 3.0 (row effect in Table 2) was 8% in accuracy, whereas the model predicts 6%. The average difference between a mean difference of 3.0 and 2.0 (column effect in Table 2) was 17%, whereas the model predicts 14%. Thus, the model captures the relative magnitude of the two effects, although slightly underpredicting them.

Busemeyer and Myung (1992) showed that their adaptive network model also produced the set of results they obtained. Actually, their model (a generalization of a class of learning models by Gluck & Bower, 1988, who derived the models from the learning rule of Rescorla & Wagner, 1972) has a similar mathematical form to the current stage of the reinforcement-learning mechanism. However, the current reinforcement-learning mechanism is simpler than their adaptive network model, and as we will show shortly, our mechanism has other attractive properties that their model does not have.

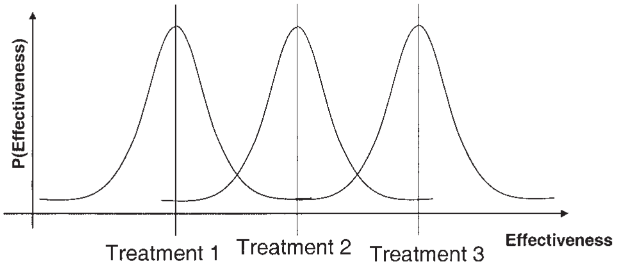


Figure 5. Distribution of effectiveness of different treatments in the experiment by Busemeyer and Myung (1992).

Table 2
Observed and Predicted Choice Proportions of the Optimal Treatment From the Experiment by Busemeyer and Myung (1992)

	Mean difference (*d*)		
SD	2.0	2.5	3.0
3.0	0.69 (0.71)	0.84 (0.80)	0.85 (0.85)
4.5	0.69 (0.69)	0.72 (0.77)	0.84 (0.83)
6.0	0.65 (0.66)	0.63 (0.73)	0.86 (0.79)

Note. Predicted scores are in parentheses.

Choosing to Gamble

In many situations one can be faced with the choice between accepting a sure gain or loss and choosing to gamble to try to obtain something better than the sure gain or to avoid the sure loss. The final experiment that we consider in this section is a relatively complicated experiment by Myers et al. (1965) that investigated such a situation. In their experiment, participants were asked to choose between a sure thing (either a gain or a loss of 1 point in different conditions) or a gamble that would return more points if successful but result in losing more points if unsuccessful. Half of the participants were assigned to the low-risk group (a range of outcomes for the risky option from $+2$ to $+6$ for a win and -2 to -6 for a loss) and the other half to the high-risk group (from ± 12 to ± 16). The researchers also manipulated the probability that the risky option would succeed.

There are two aspects to their experiment. The first is illustrated in Figure 6, which displays the learning effects, averaging over the magnitudes of payoffs. The two main findings were that (a) the choice to gamble increased when the probability of succeeding in the gamble was greater (.80 vs. .50 vs. .20) and (b) the choice to gamble was greater when the sure outcome had a negative payoff (-1 point vs. $+1$ point).[7] It is also worth noting that throughout the experiment, there was a slight overall bias to choose to gamble over the fixed alternative: Gambles were chosen on 55% of the trials, averaging over the whole experiment and all conditions.

The second effect in this experiment (and the one that has drawn more attention) concerns a pattern that appears most clearly in the asymptotic performance in the last 50 trials of the experiment. It is illustrated in Table 3. In the presence of a sure gain, participants were more likely to gamble when the magnitude of the gamble was large. In the presence of a sure loss, in contrast, participants were more likely to gamble when the magnitude of the gamble was small. This effect is a violation of a property of the classical utility theory called *independence between alternatives* (Tversky &

[7] It should be pointed out that the probability of success of gamble could depend on what the certain outcome was. There were three groups: For one group that probability was .50 independent of the certain outcome; for a second group it was .80 in the presence of a certain reward and .20 with a certain loss; and for the third group this was reversed. Thus, all groups had on average a .50 probability of success, but the last two groups had to learn the relationship between the sure payoff and the probability of success of a gamble.

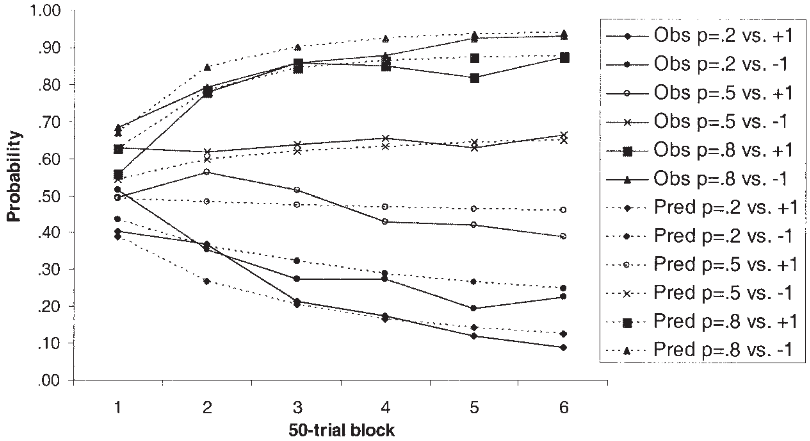


Figure 6. Proportion of gambles taken by participants and the model in each 50-trial block in the experiment by Myers et al. (1965). Obs = observed; pred = predicted.

Russo, 1969). The property of independence states that if A (gamble with large magnitude) is chosen more often in the choice set containing A and C (the sure gain) than B (gamble with small magnitude) is chosen in the choice set containing B and C, then A should be chosen more often in the choice set containing A and D (the sure loss) than B is chosen in the choice set containing B and D. This property is true of our model, because if A is chosen more often in the presence of C than B is chosen in the presence of C, then A must have greater value than B, and this greater value should also lead to more choices in the presence of D. Nonetheless, as can be seen, our model predicts this effect in this experiment, and we explain why after describing the model.

To model the data from Myers et al. (1965), we assumed four choice productions. In the presence of a sure loss of 1 point, there was one pair of productions that chose the loss or the gamble. In the presence of a sure gain of 1 point, there was another pair of productions that chose the gain or the gamble. As noted in footnote 5, separate productions were required by the fact that the gambles paid off differently in the presence of a sure gain versus a sure loss. Another critical feature of this study was that after the choice, participants were always told what they would have received if the other alternative had been chosen, which is not the case for most two-choice experiments. Because of this unique feature of the task, we, as with other models of this experiment (e.g., Busemeyer & Townsend, 1993; Gonzalez-Vallejo, 2002), assumed that the reaction to a gamble depended on both outcomes. This is consistent with results from the measurement of event-related brain potential in a similar gambling task (Yeung & Sanfey, 2004), in which participants were shown the outcome of the alternative they chose and the outcome of the alternative they did not choose. Two seemingly separate brain signals were recorded—one signal correlated with the valence of the reward and the other correlated with the magnitude of the reward (or loss). Specifically, these two signals together carried information about whether points were gained or lost *relative to* the outcomes of the alternative not chosen.

Table 3
Observed and Predicted Choice Proportions at Asymptote From Myers, Suydam, and Gambino (1965)

Probability	Magnitude of gamble	Certain payoff	Data	Model
.20	low	+1	.08	.12
.20	high	+1	.10	.13
.20	low	−1	.29	.34
.20	high	−1	.16	.17
.50	low	+1	.35	.40
.50	high	+1	.43	.47
.50	low	−1	.74	.70
.50	high	−1	.59	.54
.80	low	+1	.83	.82
.80	high	+1	.92	.89
.80	low	−1	.98	.95
.80	high	−1	.88	.92

We therefore decided to use four different parameters to represent these signals. Specifically, we assumed that if a small gamble was unsuccessful in the presence of a sure loss, this resulted in only a small regret. On the other hand, if the unsuccessful gamble was large or the alternative was a sure gain, we assumed that there was a large regret. Analogously, we assumed that if a small gamble was successful in the presence of a sure gain, this produced only a small elation, but that there was a large elation if the successful gamble was large or there had been a sure loss. The values we estimated were −2.0 for small regret and −3.3 for large regret and 2.7 for small elation and 3.3 for large elation. In principle, we could have estimated eight different utilities for all possible outcomes of the gamble (win vs. loss crossed with large vs. small crossed with sure gain vs. sure loss), but it seemed that these four parameters were adequate. In addition we estimated a 0 value for the sure gain of 1 point and −0.3 value for the sure loss of 1 point. The model does a good job of fitting the data. Looking at all 72 data points—the 36 in Figure 6 for both large and small gambles—the R^2 was .946. The standard error of the model's estimate was .060.

In particular, the model captured the violation of independence between alternatives. This depended on assessing the outcome of a gamble differently, depending on what the certain outcome was. This is not really an assumption of the model but simply borrowed from other models of the phenomenon. However, it is noteworthy that the model can predict the result of this experiment even though its basic choice rule obeys the property of independence between alternatives. The key to producing the result is that it was not the same rules (or indeed the same participants) that were choosing a gamble in the presence of a certain gain versus a certain loss or when the gamble was small versus when it was large. Therefore, the rules could evolve to have different values depending on how the outcomes were assessed. There may well be violations of the property of independence between alternatives where our explanation for the Myers et al. (1965) experiment is no longer valid, such as the choice among single gambles where probabilities are explicitly given and there is not a learning component (but see Barron & Erev, 2003). However, we would argue that these are perhaps specific to more deliberative decision making, not the kind of experience-based decision making on which we focus. Again, we should note that despite the fact that there are a number of models (e.g., Busemeyer & Townsend's, 1993, decision field theory and Gonzalez-Vallejo's, 2002, proportional difference model) for the asymptotic behavior in the Myers et al. experiment, our model is the only one, to our knowledge, that predicts the learning trends.

Summary of Results

We have tested the mechanism against four sets of data showing how preferences change over time with different experienced probabilities, magnitudes, and variabilities of payoffs. In the experiment by Friedman et al. (1964), response proportions of both the model and the participants were approximately consistent with probability matching. The simulation of this experiment established that the model showed the same sensitivities to the various probabilities of outcomes and learned approximately at the same rates as the participants. When fitting the learning data from Myers and Suydam (1964), only the reward parameters were changed and the model exhibited the same interactions between the effect of rewards and losses. Specifically, the model showed a larger effect when the loss was increased when the reward was low than when the reward was high. These two data sets showed the major effects of probabilities and magnitudes of reward in the recurrent choice literature, and our model has done a good job accounting for these effects. Although there were many published data sets that demonstrated these effects, we chose these two mainly because they presented both learning and asymptotic performance over a large number of trials.

We also showed that our model fits more complex data sets that other recent models have attempted to fit. The fit to the results from Busemeyer and Myung (1992) shows the flexibility of the model to learn from a continuous distribution of reinforcement. The fit also demonstrates the model's appealing characteristic that the feedback used for learning takes the form of a scalar error signal—that is, no explicit teacher who offers the correct answer is required. In fact, we showed that the model adapts to both the means and the variabilities of rewards based on the scalar error signal. Last, the model exhibits the choice between certain and uncertain outcomes in Myers et al. (1965). It also describes the risk-taking behavior that violated the independence axiom derived from the classic subjective utility theory. Both data sets were modeled by others, and we showed that our model is comparable to the fits of these existing models.

Part 2: Tests of Temporal Discounting

The examples so far establish that the proposed learning mechanisms can account for the choice behavior of individuals faced with alternatives that have different values and probabilities. Although we have used the full model, these examples have not really tested the assumptions in the model about delay of reinforcement. They basically are tests of the simple integrator model described in the introduction. There are two assumptions, unique to reinforcement learning, that we test in the next two sections. The one we test in this section concerns how value is discounted with delay. The next section is concerned with how the value of a reward links back over time to the actions that led to it. Much of this research has focused on animal learning, and we begin with a pair of experiments on pigeon choice behavior when the values and delay of reinforcement of two options are manipulated.

With respect to temporal discounting, we should note that some sort of discounting of future rewards is required in order for the model to converge. The typical reinforcement algorithm uses an exponential discounting, largely because of its tractable mathematical properties. However, as we will show, this seems incompatible with the actual discounting that is observed, and as others have proposed, something like our hyperbolic function is required (but see footnote 4).

Preference Reversal and Delayed Reinforcement

In the experiment by Ainslie and Herrnstein (1981), pigeons were put in a chamber containing two keys. Food was delivered with a delay after either key was pecked. However, the delay for one key was D seconds, and the delay for the other key was D + 4. In addition, the amount of food delivered for the long-delay key was always twice that for the short-delay key. For half of the pigeons, D was increased from 0.01 s to 2, 4, 6, 8, and 12 s, and then returned to 0.01 s. For the other half, D was increased from 0.01 s to 12 s, then decreased to 8, 6, 4, 2, and 0.01 s. Regardless of the delay between pecking and food, the total length of a trial was 60 s. Pigeons were kept in a constant delay condition for many sessions of 40–45 days until stable performance developed (the total experiment lasted 320 days). The results, illustrated in Figure 7, showed that subjects initially strongly preferred the smaller reinforcer at 0.01 s rather than the larger reinforcer at 4.01 s. However, as D was increased, all subjects reversed preference, choosing the larger-later option more often than the smaller-sooner option.

Figure 7 also shows the match of the model we developed, whose structure is illustrated in Figure 8. It is very similar to the choice models we used in the previous section except that we have different *Finish* productions to represent consummation of the two different rewards at different delays (in contrast, in previous models the reward was given immediately with the choice production). Because Ainslie and Herrnstein (1981) were concerned with asymptotic behavior rather than learning, we calculated the steady-state values of the system at which the values are no longer changing. We used the same parameters as in the previous models except for the rewards (i.e., r), which were estimated to be 27.6 for the small reward and 42.8 for the large reward. To help explain the behavior of the model, we consider its behavior for the condition where the two delays are 4 and 8 s. The following equations express the steady-state values for the *Prepare* production (P), the

Choose Short production that chooses the short delay (CS), the *Choose Long* production that chooses the long delay (CL), the *Finish Short* production that finishes the short delay with the reward (FS), and the *Finish Long* production that finishes the long delay with the reward (FL):

$$V_P = d(1)^*[p(\mathrm{S})^*V_{CS} + p(\mathrm{L})^*V_{CL}] = 11.52$$

$$V_{CS} = d(4)^*V_{FS} = 14.19$$

$$V_{CL} = d(8)^*V_{FL} = 14.55$$

$$V_{FS} = 27.6 + d(55)^*V_P = 28.38$$

$$V_{FL} = 42.8 + d(51)^*V_P = 46.64.$$

In the above, $d(t) = 1/(1 * kt)$ ($k = .25$) is the temporal discount,[8] p(S) is the probability of choosing the short interval, and p(L) is the probability of choosing the long delay. With the noise parameter set at the fixed value of 1.0, the value of $p(\mathrm{S}) = .412$ and $p(\mathrm{L}) = .588$ according to the conflict-resolution equation. To reach these steady-state utilities, one has to guarantee that each choice gets sampled enough, and therefore it is necessary to use a noisy conflict-resolution equation so that each item has a minimum probability of being selected. Without this, the model can sometimes reach asymptote on the preferred option before the other option and come to select it exclusively, blocking further learning on the other option. This is consistent with the procedure in many animal studies (e.g., Mazur, 1985), in which forced-choice trials are interspersed with free-choice trials to ensure the experience with each alternative.

The indifference point is where the values of CS and CL are equal, which is where the ratio of $d(t)/d(t + 4)$ is the same as the ratio of V_{FL}/V_{FS}, and that equals 1.55. With $k = .25$ and solving for $d(t)/d(t + 4) = 1.55$, we get $t = 3.3$ s. With respect to the fit of the model, we obtained a fit of $R^2 = .956$ and a standard error of the estimate of .068. This experiment is just one of the many studies in the animal conditioning literature that justifies a temporal discount function such as the one in our model; on the other hand, it is the first that shows the dramatic preference reversal as the delays of reinforcement were manipulated. We will show shortly why the exponential discount function had trouble producing the preference reversal.

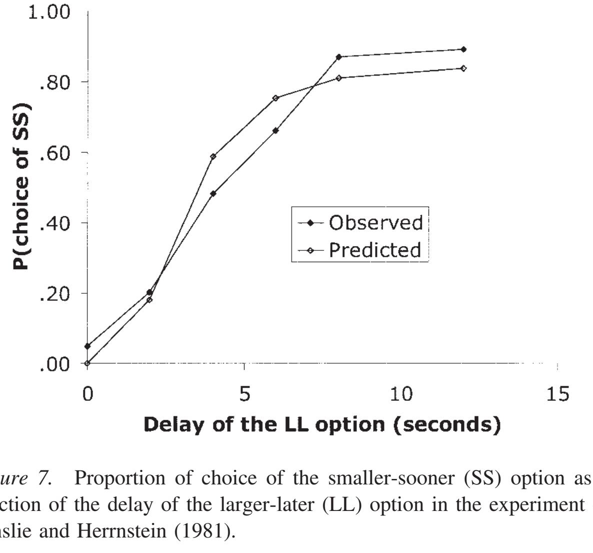


Figure 7. Proportion of choice of the smaller-sooner (SS) option as a function of the delay of the larger-later (LL) option in the experiment of Ainslie and Herrnstein (1981).

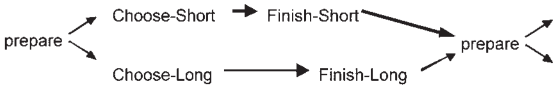


Figure 8. The model for the experiment by Ainslie and Herrnstein (1981).

Reinforcement Delay–Amount Trade-Offs

Although the previous experiment justifies a temporal discount, it does not justify the hyperbolic function. Evidence for the hyperbolic form comes from the experiment of Mazur (1985), which introduced an adjusting procedure to find the indifference points for various sets of delayed reinforcement. In an adjusting procedure, subjects choose between a standard alternative and an adjusting alternative. In Mazur's experiment, the standard alternative delivered pigeons a certain amount of food (R_F) after a fixed delay D_F. The delay for the standard alternative was different in each of the nine conditions but was fixed throughout a condition. The adjusting alternative delivered a larger amount of food (R_A) after an adjusting delay (D_A). The amount of food for the adjusting alternative was three times that for the standard alternative. (The delay between trials was 15 s, independent of choice.) If a pigeon chose the adjusting alternative on two consecutive trials, D_A was increased by 1 s; if the standard alternative was chosen on two consecutive trials, D_A was decreased by 1 s. When the pigeon chose the two alternatives about equally often, the adjusting delay could be considered the indifference point. At the indifference point, the values of the CS and CL productions are equal. In each of the nine conditions, pigeons were given a minimum of 12 64-trial sessions, until the choice stabilized. The standard delays for the nine conditions were $D_F = 0, 1, 2, 6, 6, 10, 12, 14,$ and 20. Figure 9 shows the mean indifference points for each standard delay. The best fitting regression line indicates a slope of 2.4 and a y-intercept of 2.2.

At the indifference point the values of the two choices will be the same, and we can use an extension of the asymptotic equations given earlier to find two expressions for this value corresponding to its calculation at the long and short delay:

$$V = d(\mathrm{S})[R_S + d(15)d(1)V],$$

$$V = d(\mathrm{L})[R_L + d(15)d(1)V].$$

These equations and the definition of the discount function $d(t) = 1/(1 + kt)$ can be manipulated to give an expression for the long delay (L) that matches a particular short delay (S):

$$L = R_L/R_S{}^*S + \frac{(R_L/R_S\ -\ 1)[1 - d(15)d(1)]}{k}.$$

Thus, the calculation of the matching delay depends on the ratio of the larger to the smaller reward (R_L/R_S) and not on their actual

[8] Note that because each trial took 60 s, the time to the next trial was slightly less in the case of the long interval. For instance, in the case of a 4-s short delay versus an 8-s long delay it would be 55 s after reward to the next trial for the short delay and 51 s for the long delay.

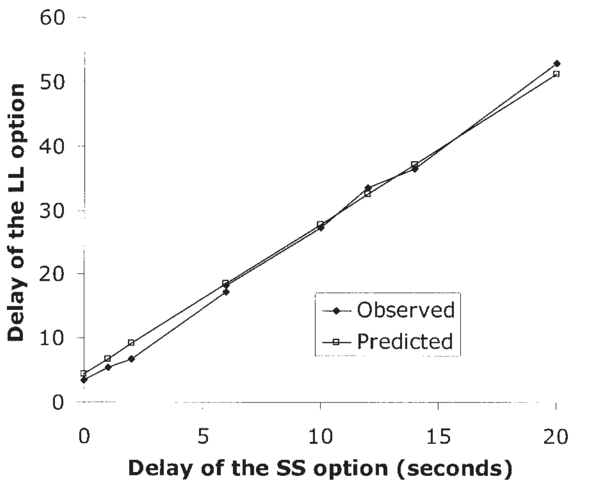


Figure 9. Indifference points of the two delayed rewards in Mazur's (1985) experiment. LL = larger-later option; SS = smaller-sooner option.

values. The best fitting value of this ratio is 2.34. With this value of the ratio of the rewards, the model predicts a linear equation with this ratio as the slope and the intercept of 4.46 s. The model fits the data well, $R^2 = .998$, and standard error in prediction was 1.28 s, which is quite adequate given that the standard deviation in the empirical values is 1.80 s. This is striking support for our theory, particularly for the underlying discount equation. It predicts the linear relationship between the length of the two intervals, and it predicts both the slope and the intercept of that equation with an estimate of a single parameter (the ratio of the rewards).

As we noted, the most typical temporal discount function in existing reinforcement-learning models is the exponential function. We can substitute the exponential function $d(t) = a^t$ into the equations above for value in the short and long delays to calculate the relationship between the two at the point of equilibrium:

$$L = S - \frac{\ln[R_L/R_S] + \ln[1 - a^{16+S}(1 - R_S/R_L)]}{\ln(a)}.$$

This shows that the relationship is again predicted to be linear but this time with a slope of 1 and an intercept that depends on the ratio of the large to the small reward.[9] This equation shows that the fundamental problem with an exponential discount function is that it predicts an additive shift in the matching delay for a larger reward. However, we showed that the relationship between the two delays is basically multiplicative. We should point out that the exponential discount function predicts that the two curves will never cross over. Thus, the exponential function is not capable of producing the preference reversal as shown by Ainslie and Hernstein (1981).

Summary of Results

We showed that the model fit the two sets of data well. In the experiment by Ainslie and Herrnstein (1981), both the model and the subjects reversed preference for the larger-later rewards instead of the smaller-sooner rewards as the delay increased. In the adjusting schedules of the experiment by Mazur (1985), the model stabilized at the same indifference points for the various conditions as the subjects did. Although the same hyperbolic function was used by Mazur, our model was at a lower level than the model by Mazur, which aimed at predicting the stabilized indifference points. Unfortunately we do not have the learning data for Ainslie and Herrnstein or Mazur's experiments, but our model is potentially able to match the learning data that the molar model of Mazur is not capable of doing.

Part 3: Tests of Credit Propagation in the Acquisition of Action Sequences

One of the powerful features of the TD algorithm is its ability to propagate credit back to previous productions. However, this process takes time. Initially, only the production that directly leads to the reward gets credit. In the next cycle of production firings, some of that credit propagates back to the previous production. Eventually, credit can find its way back to a critical early production through a chain of productions that leads to the reward. This credit propagation mechanism is essential for learning action sequences, as credits are assigned to all productions that eventually lead to a reward, including the critical production that initiates the sequence of productions. Strong evidence for such propagation of credit would seem to be found in the research on goal gradients in maze learning. We describe a model for the classic experiment by Tolman and Honzik (1930) and then point out the problem with this experiment and similar experiments. Then we describe the experiment we performed to obtain more definitive data on this issue.

The Maze Experiment by Tolman and Honzik With Rats

Tolman and Honzik (1930) used a 14-unit T maze (see Figure 10) to study how rats learn to eliminate blinds. After preliminary training on how to manipulate the gates and curtains at the junction points, each of the 36 rats was given one run a day in the maze for 17 days. Time and error records were kept, but as we are not predicting time here, we focus on the error data. An entrance into a blind was considered an error. Gates prevented returns into units just visited, and so there was not the potential to retract successful choices and go backward. Figure 11a shows the learning measured as total number of errors made by the 36 rats, and Figure 11b shows the proportion of errors to each of the blinds in Figure 10. Figure 11 shows that rats learned to reduce the number of errors with practice but that errors were not distributed equally for all blind positions. Overall, there was a trend for fewer errors closer to the food, but it is clear that other factors were at work as well. In addition to proximity to the food, fewer errors were made when the correct choice was to go to the right or to go down. Overall, more correct moves are to the right, and so moving right is a good guess. The advantage of the downward moves is a second-order goal-gradient effect in that correct moves close to the food are down whereas those farther away are up. Although there are these important directional effects, the overall pattern of results is often taken as evidence supporting the notion of a goal gradient: The influence of the final reward is graded according to how far away the stimulus is from the goal (or in this case, the food reward).

[9] Although the intercept is expressed above as a subtraction, it is a positive intercept because $\ln(a) < 0$ and the ratio of the rewards is greater than 1.

Figure 11 also displays the predictions of our model for this task. The model assumes that there are two productions specific to each choice point, each requesting a move in the two possible directions. In addition, there are four general directional productions, *move down*, *move up*, *move left*, and *move right*. Given the design of the maze, at any choice point, only two of the general directional productions are applicable. The first set of choice-specific productions would eventually learn the maze, but the second set allowed for directional biases, particularly in the early stages of learning. One production fired at each choice point. If it led to a blind, it was punished with a negative value. If it did not lead to a blind, nothing happened, except for the last production, which would be rewarded positively with the food. Eventually, the results from the correct final choice can propagate back to the earlier choices. The two parameters estimated were the positive utility of the food (estimated at 120) and the negative utility of going down a wrong choice and having to back up (estimated at −19). All other parameters were the same as in the other models. Figure 11 displays the correspondence between this model and the data, and it is quite striking. For the learning curve, R^2 is .973 and mean error is 13. For the blind performance, R^2 is .718 and mean error is 2.2%.

A Maze Experiment With Humans

After working with the Tolman and Honzik (1930) data set, we came to the conclusion that it had three unfortunate aspects with respect to assessing credit assignment in the TD algorithm. First, the directional effects have a major impact on the goal gradient, as is apparent from Figure 11b. Although we captured much of this effect with our use of left, right, up, and down productions, there may have been other directional biases to complicate matters (rats are particularly good with respect to directional sensitivity; see O'Keefe & Nadel, 1978). Second, rats came immediately to a blind after a mistake and were forced onto the right path. Thus, the blinds provided immediate feedback, and it was not necessary for credit to propagate back from the goal. We modeled this by having a negative value for a wrong choice. Indeed, it was necessary to have these punishment factors because of the third problem with the Tolman and Honzik experiment: that it was simply not possible for credit to effectively work its way all the way back through 14 choices from the goal in 17 trials.[10] To address these problems, we chose to run an isomorph of a maze experiment with humans. The character of the isomorph was designed to eliminate any complexities produced by spatial reasoning. However, it retained the characteristic that learning requires the acquisition of action sequences that depends critically on the assignment of credit to the right actions executed at different points in time.

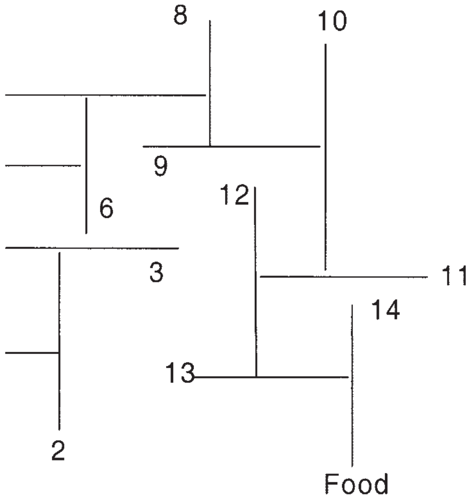


Figure 10. T maze used in Tolman and Honzik (1930). The numbers represent blinds in the maze. When a rat chose the wrong direction and entered a blind, it needed to turn around and go in a different direction to go to the next T junction.

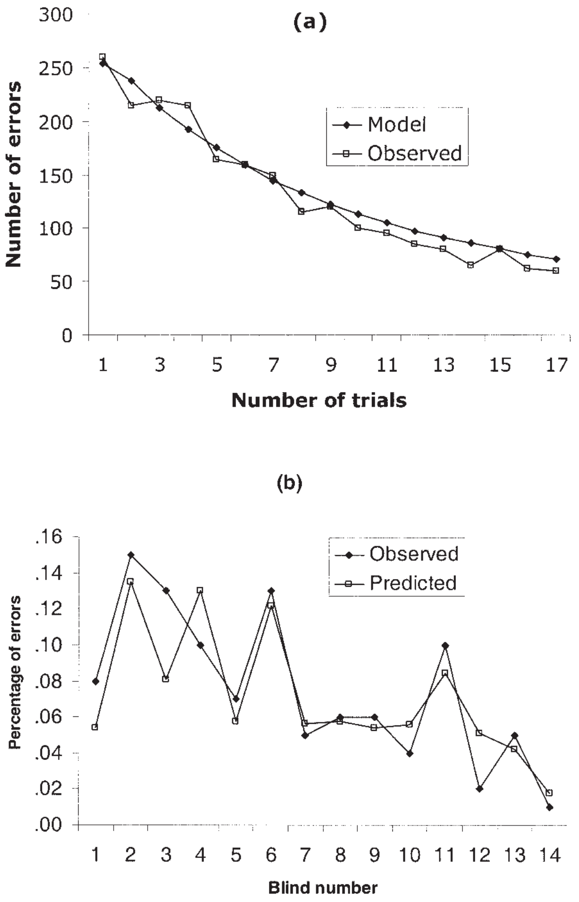


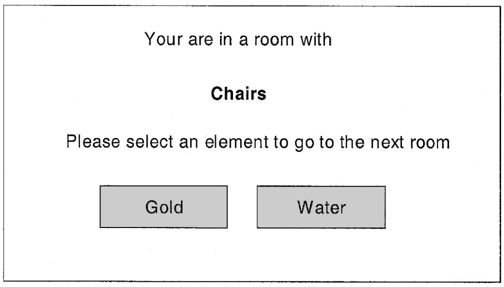
Figure 11. (a) Observed and predicted number of errors and (b) percentage of errors in each of the blinds in the maze experiment of Tolman and Honzik (1930).

The maze-searching task. The maze-searching experiment was designed to directly test the goal-gradient hypothesis as suggested by the results from the maze experiment by Tolman and Honzik (1930). To eliminate the complexities of spatial reasoning, we created an artificial maze-searching game, presented on a

[10] While one might think one choice point could be learned per trial and so 17 trials would be enough for 14 choices, it takes a number of trials for the assessment to build up at a choice point before any significant credit can propagate back to the next choice point.

computer, in which participants were asked to make arbitrary association of objects to guide their choices as they progressed through a series of simulated rooms. The screen symbolizing each room contained a single object term (e.g., *chairs*, *phones*)[11] and terms for two elements (chosen from four—*gold*, *wood*, *water*, and *fire*—intended to be analogous to the four directions in the maze experiment). Figure 12 shows a screenshot of a room. By selecting an element, participants were taken to another room with a new object and another two elements. The participant's task was to find the correct sequence of elements that would lead through the correct sequence of rooms to finish the trial. The correct element depended on the object in the room. Thus, a particular element was successful only in rooms that contained a particular object. Performance on the task thus depended on learning which element to select in the presence of each object. This kind of learning is central to many skill-learning situations, as the main component of a skill is to know when to apply the right action given a particular context of cues in the environment. The learning of the association of the object and right element (i.e., actions) can therefore be considered one of the core learning components in skill acquisition.

The crucial aspect of the task is that feedback was given only after three correct choices had been made or when a dead end was reached. Figure 13 shows the map of the rooms (which was not shown to participants). In each of the rooms, there were four possible object–element triplets. When participants reached a room, one of the object–element triplets was randomly sampled and presented on the screen. The participant had to choose between the two elements, and only one would lead to the correct next room (provided the participant was still on the correct path). All trials began in Room 1. When three correct choices were made (i.e., when participants progressed from Room 1 to Room 2 to Room 3 to "finish" in Figure 13), the trial would end and 3 points would be given. Otherwise, if one or more of the choices were wrong, participants would reach one of the dead ends (i.e., D1 to D7 in Figure 13) and 1 point would be deducted from the total score. When a dead end was reached, participants had to reset (by clicking a button in a pop-up window) and try again. After the reset, participants would be taken to the earliest room where they had made the wrong choice.[12] For example, if the participant correctly chose Room 2 in Figure 13 but then erroneously chose Room 4 followed by D3 and reset, the participant would be taken to Room 2. Another object–element triplet was then sampled and presented to the participants in Room 2, and the game continued until they finished the trial. The fact that the object–element triplet was resampled meant that it was very difficult for participants to determine what room they had been sent back to, and in effect the only direct feedback they received was on moves that directly led to a dead end (and they lost a point) or when they reached the finish room (and they gained 3 points).



Your are in a room with

Chairs

Please select an element to go to the next room

Figure 12. A screenshot of an example room in the maze-searching task.

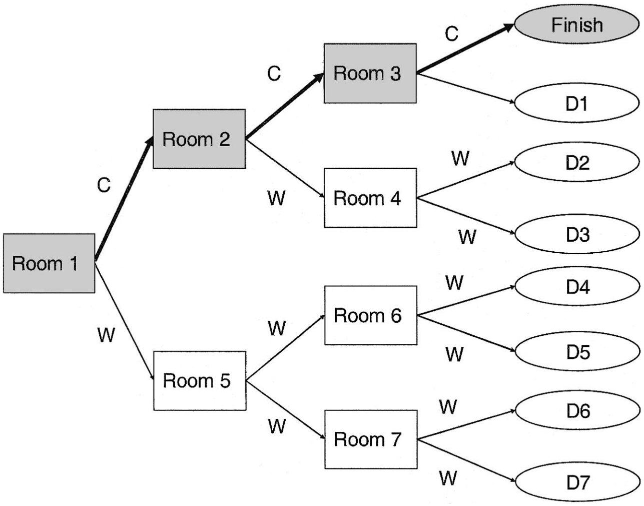


Figure 13. The structure of the seven rooms in the maze-searching task. When a dead end (D1, D2, D3, D4, D5, D6, or D7) was reached, participants needed to reset. C = correct choice; W = wrong choice.

There were 12 correct object–element associations (4 in each of the 3 correct rooms on the way to finishing the trial). Because feedback was given only when the correct path was found or when a dead end was reached, learning of the correct early object–element associations required propagating credit back from later object–element associations. If the goal-gradient hypothesis is correct, object–element associations closer to the feedback should be learned faster than those farther away from the feedback (i.e., learning should be fastest in Room 3, followed by Room 2, and slowest in Room 1).

Note that this is a problem that is not naturally represented in terms of choice of states, which is the more traditional use of reinforcement learning in artificial intelligence. Participants are making a choice of an operator, not a unique state; when they choose an operator, they can transit to any of the four different states depending on what operator comes next. It is also the case that there are $4 \times 4 \times 4 = 64$ correct operator sequences depending on what states the participant transitions to. Therefore, it is not

[11] The objects used were *torches*, *spiders*, *tables*, *fishes*, *dishes*, *books*, *computers*, *cigarettes*, *smoke detectors*, *televisions*, *radios*, *pencils*, *wallets*, *keys*, *rats*, *cats*, *chairs*, *telephones*, *bags*, *cups*, *glasses*, *watches*, *cans*, *folders*, *magazines*, *newspapers*, *envelopes*, and *stamps*.

[12] There were a number of reasons for not taking participants all the way back to the beginning each time. One was that we wanted to give participants one experience with the correct choice at each level on each trial. It also made the experiment a little less frustrating, because if we had always sent participants back to the beginning, they would have had a $(1/2)^3$ chance of getting it right and thus have had to make three choices for each mistake—which implies something like 24 moves to solution at the beginning before they learn. We also feared that always sending them back to the beginning might invoke some strategy of trying to get that first move right first.

feasible to learn each sequence separately. Rather, the participant must learn the 12 rules specifying what elements to select in the presence of a particular object. Also, it is not obvious what level (first, second, or third) one is at, because one is not told the level to which one is sent back. Thus, the only real feature is the presented object and not the point in the maze. Except for the last move, the only feedback one gets as to whether one has applied a correct operator is whether one is then transitioned to a state where one can apply a correct operator. So it is a task that really can only be learned by credit-assignment mechanisms like those in reinforcement learning.

Method. Twenty members of the Carnegie Mellon University community (age ranged from 19 to 32; 9 women, 11 men) participated in the experiment. Each participant finished 200 trials and was paid either $8 or 1 cent per point, whichever amount was greater. Participants spent an average of 1 hr in the game. They were instructed that through experience with the rooms they could learn which elements would be correct in the presence of different objects and that the object–element associations would stay the same throughout the experiment (the actual instructions given to the participants can be found in the Appendix). Participants were told that the only feedback they would be given was when they had made three correct choices or when they had reached a dead end. They were told that when they reached a dead end, they had to click the "reset" button, which would take them to another room. Participants were not informed of the structure of the rooms. All stimuli were presented and all responses were recorded via a standard monitor controlled by a personal computer system.

Results. Figure 14 shows the mean number of rooms visited in a trial in each of the 10 20-trial blocks. Participants visited roughly 10 rooms in the beginning trials and approached the asymptote of 4 rooms per trials after roughly 100 trials (perfect performance was 3 rooms per trial). Figure 14 also shows the predictions of our model for the task. The model represents each object–element association as a production. Similar to the model for the maze in Tolman and Honzik (1930), there are always two productions that compete in each of the rooms. However, in this task, we do not need the general directional productions to account for the directional biases observed in the Tolman and Honzik experiment. We constrained the major parameter values of the model to be the same as in previous models (i.e., delayed-reward parameter $k = .25$, learning rate $a = .05$, and noise parameter $t = 1.0$)[13] and varied only the reward parameter to fit the data. The reward parameter was estimated to be 12. Because learning was relatively insensitive to the punishment parameter we set it to be -12 (just to be simple), which is the negative of the reward parameter. We obtained a fit of $R^2 = .947$, with an average error of .54. In general, the model captures the learning curve of the participants. Given the constraints of the parameter space, we consider the model to have done a good job fitting the data.

Initially, the model randomly picks one of the productions to progress through the rooms. The rewards and punishment eventually propagate back to the earlier productions. Figure 15 shows the percentages of correct choices made in Rooms 1, 2, and 3. The main effects of rooms and trials were significant, $F(2, 57) = 6.96$, $MSE = 2.42$, $p < .01$, $\omega^2 = .074$. The differences between Room 2 and Room 1 and between Room 3 and Room 1 were significant, $F(1, 57) = 3.87$, $MSE = 1.35$, $p = .05$, $\omega^2 = .041$, and $F(1, 57) = 13.9$, $MSE = 4.84$, $p < .01$, $\omega^2 = .186$, respectively. The difference between Room 3 and Room 2 approached significance, $F(1, 57) = 3.11$, $MSE = 1.08$, $p = .08$, $\omega^2 = .030$. In general, the results are consistent with the goal-gradient hypothesis: The object–element associations closer to the reward were learned faster. Figure 15 also shows the predictions of the model. With the same constraints of parameter values, we obtained a fit of $R^2 = .857$, with an average error of .07. The model captures the goal gradient of reinforcement by learning the object–element associations in Room 3 fastest, followed by Room 2, and then Room 1. The model, however, learned slower than the participants in Room 1 during early trials. We speculate the reason is that there was some inherent saliency of Room 1 as it was the first room of each trial and perhaps some participants were setting up special strategies to try to learn this choice.

Summary of Results

The model fits the two sets of data fairly well. In both cases, the reinforcement-learning mechanism propagates discounted credit back to previous productions, which in effect produces the goal gradient—that is, learning of rewards is graded and is more efficient at choice points that are closer to the rewards. The results demonstrated that through repeated exposures to the same reward structure in the maze, the "built-in" credit-assignment mechanism in the reinforcement-learning algorithm exhibited the goal gradient found in many animal studies (see Killeen, 1994, for a review). In the second experiment, the human data we collected from the maze-searching task supported the basic premise of the goal-gradient hypothesis, and our model exhibits a similar goal gradient as a result of the credit-assignment mechanism. Specifically, the credit-assignment mechanism allows the model to learn which of the many possible object–element rules are more likely to be correct, and performance improves as these correct rules are selected more often across trials.

General Discussion

We have presented a reinforcement-learning model of recurrent choice implemented in a general production-system framework. The model was inspired by recent understandings of the dopaminergic signals in the basal ganglia and their relation to reinforcement learning. To demonstrate the value of this endeavor we have identified representative results that covered the major manipulations in the recurrent choice literature, namely, the probabilities, magnitudes, variabilities, and delay of reward. To extend the model and test its ability to account for skill learning, we performed a study that tested the goal-gradient hypothesis in a maze-searching task. We showed that the reinforcement-learning models we constructed provided general moment-to-moment predictions of the strength of preference of alternatives. Although the successful use of temporal difference methods has been demonstrated in many artificial intelligence systems, to our knowledge, no attempts have been made to test their psychological validity against the same range of empirical data as we did. The fact that the use of a fixed set of learning and delayed-reward parameters explained

[13] The time between moves was set to 3 s, which was approximately the average intermove time for the participants.

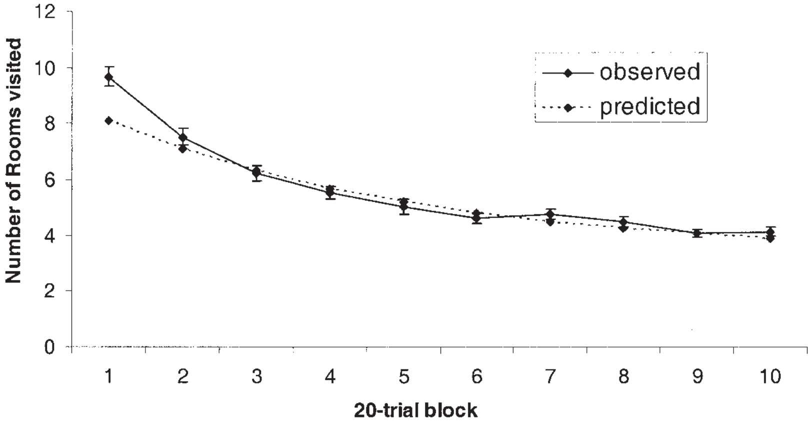


Figure 14. Observed and predicted number of rooms visited in each 20-trial block of the maze-searching task.

such a wide range of choice phenomena provides strong psychological validity to the mechanism.

The reinforcement-learning model, in its simplest form, is similar to an integrator model that accumulates information about the consequences of a choice. We showed that the model, when combined with the conflict-resolution equation, provides a stochastic–dynamic description of the recurrent choice process that lies at the heart of learning and performance in choice behavior. We showed that the reinforcement-learning model has two unique characteristics that are often neglected in other recurrent choice models: how value is discounted with delay and how the utility of a reward will link back over time to the actions that led to it. These characteristics are essential for skill learning, which often requires the learning of action sequences by delayed feedback.

Multistep Recurrent Choice and Skill Learning

We showed that the temporal discounting of rewards produced the results from the experiment of Ainslie and Herrnstein (1981), in which preferences reversed from a larger-later reward to a smaller-sooner reward as the difference of delays of the two rewards increased. Further support was provided as we showed that the model provided striking fits to the data from Mazur's (1985) experiment, in which a wide range of delays of reinforcement was manipulated. The discounting property is then combined with the credit-assignment mechanism to predict behavior during skill learning. Skill learning can often be cast as multistep recurrent choice situations, in which several actions are performed before reinforcement on the full course of actions is received. In these situations, not only are rewards temporally discounted, but rewards must propagate back to the appropriate actions that are responsible for the rewards. The reinforcement-learning model provides a straightforward explanation of how rewards propagate back to earlier actions. Initially, only the production that leads to reward gets credit. The next time, some of that credit propagates back to the previous production. Eventually, credit can find its way back to critical early productions in a long chain of productions leading to a reward. We showed that our model provides good fits to the data both from the maze-learning experiment by Tolman and Honzik (1930) and from the experiment we designed. Both data sets support the existence of a goal gradient of reinforcement—that is, credit received is graded according to how close the actions are to the reward.

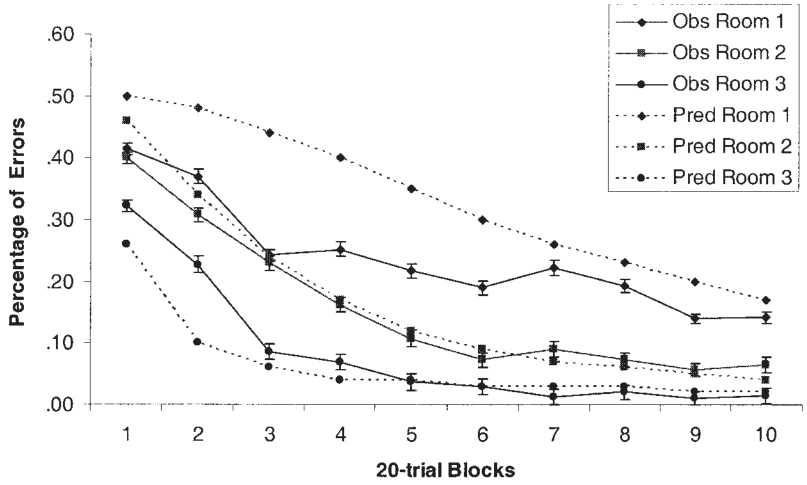


Figure 15. Observed (obs) and predicted (pred) percentages of errors in Rooms 1, 2, and 3.

In the animal literature, it has long been noted that a reinforcer's effects are not limited to just the response that immediately preceded it. Rather, the reinforcement reaches back toward earlier responses, but the effectiveness diminishes with temporal distance (see Killeen, 1994, for a review). Killeen (1994) proposed that the animal keeps tracks of past responses in short-term memory, and reinforcement will strengthen the responses in short-term memory with a decay gradient, so that the reinforcement will have less and less impact on the responses farther away from the reinforcement. Killeen's theory raises the questions of how the responses are represented in memory and, to be able to explain human behavior, whether previous responses are actually remembered and whether this is a criterion for any response to be reinforceable. Indeed, recent neuroscience studies on rats have shown that reward-related response learning does not seem to rely on declarative memory representations of past responses (e.g., Packard, 1999; Packard & McGaugh, 1996). Rather, declarative (hippocampus) and nondeclarative (striatum) memory systems seem to work independently of each other during maze learning. As previous studies have found a close match between neural activities in the striatum and the reinforcement-learning mechanism during skill learning (e.g., Knowlton, Mangels, & Squire, 1996; Poldrack, Prabhakaran, Seger, & Gabrieli, 1999; Schultz et al., 1997), our model is more closely related to the nondeclarative response-learning mechanism found in the striatum in different reward-related learning tasks. However, it seems likely that complex skill learning may involve both memory systems and their complex interactions.

The interactions of the declarative and the nondeclarative memory systems have also been shown by recent neuroimaging studies. For example, Poldrack et al. (1999), using a probabilistic classification task, showed that the striatum plays an active role during the learning process. In that task, participants were presented with two cards with different patterns on them. They were then asked to press a switch if they thought the patterns signified rain, guessing at the outset but using feedback on each trial to learn which patterns signified rain. Each pattern combination had a certain probability of signifying rain. In addition to the significantly higher activation in the striatum, Poldrack et al. (1999, 2001) found that the anterior portion of the medial temporal lobe, which is often associated with declarative memory retrieval, was deactivated compared with the baseline activation when the striatum was active. The significant negative correlation between the activities of these two memory systems suggests that they may play dissociable roles in skill learning. Knowlton et al. (1996) also found that patients with dorsolateral frontal lesions performed just as well as healthy individuals in a probability-learning task, whereas patients with Huntington disease (which compromises the function of the striatum) performed significantly worse than healthy individuals (see also Morris, Miotto, Feigenbaum, Bullock, & Polkey, 1997).

Comparisons to Other Models

A number of the data sets presented in this article have been modeled by other theories of choice. In this section we present some comparisons of our model to the existing models. Two of these theories are the decision field theory (DFT) of Busemeyer and Townsend (1993) and the proportional difference (PD) model of Gonzalez-Vallejo (2002). Although they are not restricted to models of recurrent choice, both of these models have provided good fits to the asymptotic performance of participants in the Myers and Suydam (1964) and Myers et al. (1965) studies. However, instead of focusing on how preferences change with repeated feedback on previous choices, both theories focus on the *local context* defined by the alternatives in single-choice studies. In DFT, strengths of preferences are represented by the difference of two weighted averages of the *valences* of the alternatives. In PD, strengths of preferences are represented by a mechanism that performs trade-offs between attributes within a particular representational structure defined by a function that calculates the proportional difference between the alternatives. In contrast, our model calculates moment-to-moment preferences from the histories of reinforcement of different alternatives. The appealing aspect of our model is that we are capable of explaining the learning data that DFT and PD cannot. Conversely, the focus on local context allows both DFT and PD to predict results that show violations of transitivity (e.g., Mellers, Chang, Birnbaum, & Ordonez, 1992; Tversky, 1969) that our model cannot predict, as those studies are not conducted under recurrent choice situations.

A recent recurrent choice model, called reinforcement learning among cognitive strategies (RELACS), was proposed by Erev and Barron (2003; see also Erev et al., 1999). RELACS is a learning model that captures the changes in choices as a function of the experienced payoffs. Because RELACS is derived from the same reinforcement learning algorithm as our model, it is not too surprising that the behavior of RELACS is very similar to our model. For example, Erev and Barron showed that their model described the learning trends in Myers et al. (1965) and some of the similar effects such as probability learning that we presented in Part 1 of our section *Testing the Mechanism Against Empirical Data*. In addition, there are reinforcement-learning models that show probability matching behavior (Egelman, Person, & Montague, 1998; Montague, Dayan, & Sejnowski, 1996). Our model, however, extends the scope of the mechanism to successfully describe empirical results that show how choice behavior is sensitive to delay of rewards and the sequential dependencies of choice actions in skill learning. We believe that our effort is complementary to the work by Erev and his colleagues and others.

To return to the distinction at the beginning of this article, our model is concerned with outcome of quick, nondeliberate decisions reflecting statistical learning over many experiences and not deliberative decision making. It would seem that DFT and PD are more appropriate to the latter. Nonetheless, our model and theirs have been applied to predicting the same asymptotic performance in the Myers and Suydam (1964) and Myers et al. (1965) experiments, and this raises the question of what kind of decision making participants were actually engaged in during those experiments. Undoubtedly, it is a mix of the two, but our success at predicting the learning trends leads us to believe that the behavior was dominated by the kind of statistical learning that our model addresses (see, e.g., the discussion by Estes, 2002).

It is interesting to compare our model with the adaptive network model (a generalization of a class of learning models by Gluck & Bower, 1988) that produces the set of results from the experiment of Busemeyer and Myung (1992). We find that the mathematical form of the adaptive network model is similar to the general form of our model (without the temporal discounting of rewards). Their model has been applied to learning the relevance of multiple cues in a categorization task. Anderson and Matessa (1998) showed that the kind of utility learning in ACT–R can predict, at least qualitatively, the results from Gluck and Bower's complex categorization experiments on cue learning.

The good fits to the data from the experiments by Ainslie and Herrnstein (1981) and Mazur (1985) show that the hyperbolic discount function in the reinforcement-learning model provides a good description of how delayed rewards are discounted. Although the same hyperbolic discount function has been used by Mazur and others (e.g., Loewenstein & Prelec, 1991), the current form of our model predicts learning data that cannot be predicted by the steady-state model of Mazur or Loewenstein and Prelec.

The data from Tolman and Honzik (1930) and the maze-searching task have provided further support to the temporal discounting function. In addition, the good fits of the model to the maze-searching task data show that the temporal discounting function provides good description to the learning process. The good fits also show that the credit-assignment mechanism in the model explains how people learn to choose a sequence of actions that eventually leads to the rewards. Although many machine learning methods and their properties have been presented (see, e.g., Sutton & Barto, 1998) to show how an artificial agent learns to choose among sequences of actions from delayed feedback, no attempt has been made to study how people learn to assign credit to different actions in these situations.

The cumulative effects model by Davis, Staddon, Machado, and Palmer (1993) addresses animal choice behavior in situations involving extinction of reinforcement. In one of these situations (Davis & Staddon, 1990), the animal first received reinforcement from the right alternative for n sessions, then received reinforcement from the left alternative for m (where $m < n$) sessions, and then no reinforcement was given (extinction). During extinction, the animal began the first session with an almost exclusive left preference, but preference shifted through indifference to a right preference by the fourth session of extinction. Not only does the cumulative effects model predict the spontaneous recovery of an earlier preference (i.e., regression) in extinction, it also predicts phenomena such as the dependence of learning rate on the frequency of reversals of reinforcement and improvement in learning speed across reversals of reinforcement, all of which require some "memories" of the past histories of reinforcement. Davis et al. showed that a simple integrator model, without separately keeping track of past histories of reinforcement, fails to predict these phenomena. Because our model, in its general form, is an integrator model, changes need to be made if the model is to produce the same behavior as the cumulative effects model predicts. Although it is interesting and important to implement in our model, we have decided not to put this possible extension in this article to avoid overly complicating the basic model.

Advantages of Implementing the Model in a Production-System Framework

By implementing the learning mechanism under a general production-system framework, successful models were produced and reviewed in this article. Our intention is that the model can be incorporated into a larger cognitive architecture such as ACT–R, Soar, or Epic. We believe that by putting the choice mechanism into a larger architecture, one can apply the same ideas to the many aspects of recurrent choice behavior that occur in complex tasks, such as sequence learning, strategy selection, and problem-solving search. This allows the construction of choice models in more real-world, complex situations (e.g., the dynamic tasks facing anti-air warfare coordinators). In complex, dynamic tasks (Anderson et al., 2004; Fu et al., 2004), a model of choice often requires the integration and orchestration of several cognitive mechanisms, and each of these is driven by simple choices. For instance, choices are required about what particular memory elements are needed at a particular time, what part of the visual array to attend to, or which method to apply to make an edit to a text. We believe that the advantage of the integration provided by a cognitive architecture is the potential to apply the same insights to all of these decisions.

References

Ackley, D. H., Hinton, G. E., & Sejnowsky, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9,* 147–169.

Ainslie, G., & Herrnstein, R. J. (1981). Preference reversal and delayed reinforcement. *Animal Learning & Behavior, 9,* 476–482.

Alexander, G. E., Crutcher, M. D., & Delong, M. R. (1990). Basal ganglia–thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Progress in Brain Research, 85,* 119–146.

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience, 12,* 505–519.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science, 29,* 313–342.

Anderson, J. R., Bothell, D., Byrne, M., Douglass, D., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111,* 1036–1060.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, J. R., & Matessa, M. (1998). The rational analysis of categorization and the ACT–R architecture. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 197–217). Oxford, England: Oxford University Press.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16,* 215–233.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron, 30,* 619–639.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121,* 177–194.

Busemeyer, J. R., & Townsend, J. (1993). Decision field theory: A dynamic–cognitive approach to decision-making in an uncertain environment. *Psychological Review, 100,* 432–459.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning.* New York: Wiley.

Davis, D. G. S., & Staddon, J. E. R. (1990). Memory for reward in probabilistic choice: Markovian and non-Markovian properties. *Behaviour, 114,* 37–64.

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

Delgado, M. R., Locke, H. M., Stenger, V. A., & Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive, Affective, and Behavioral Neuroscience, 3,* 27–38.

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation, 14,* 1347–1369.

Egelman, D. M., Person, C., & Montague, P. R. (1998). A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience, 10,* 623–630.

Erev, I., & Barron, G. (2003). *On adaptation, maximization, and reinforcement learning among cognitive strategies*. Working paper, Technion Institute of Technology, Haifa, Israel.

Erev, I., Bereby-Meyer, Y., & Roth, A. E. (1999). The effect of adding a constant to all payoffs: Experimental investigation, and implications for reinforcement learning models. *Journal of Economic Behavior and Organizations, 39,* 111–128.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review, 9,* 3–25.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003, March 21). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299,* 1898–1902.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2005). Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. *Behavioral and Brain Functions, 1,* 7.

Friedman, M. P., Burke, C. J., Cole, M., Keller, L., Millward, R. B., & Estes, W. K. (1964). Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 250–316). Stanford, CA: Stanford University Press.

Fu, W.-T., & Anderson, J. R. (2004). Extending the computational abilities of the existing procedural learning mechanism. In D. F. Kenneth, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 416–421). Mahwah, NJ: Erlbaum.

Fu, W.-T., Bothell, D., Douglass, S., Haimson, C., Sohn, M.-H., & Anderson, J. A. (2004). Learning from real-time over-the-shoulder instructions in a dynamic task. In *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 100–105). Mahwah, NJ: Erlbaum.

Fu, W.-T. & Pirolli, P. (in press). SNIF-ACT: A model of user navigation on the World Wide Web. *Human-Computer Interaction.*

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Gonzalez-Vallejo, C. (2002). Making tradeoffs: A probabilistic and context-sensitive model of choice behavior. *Psychological Review, 109,* 137–155.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (in press). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review.*

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

Hinton, G. E., & Sejnowsky, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations* (pp. 282–317). Cambridge, MA: MIT Press.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109,* 679–709.

Houk, J. C. (1995). Information processing in modular circuits linking basal ganglia and cerebral cortex. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 3–9). Cambridge, MA: MIT Press.

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. 233–248). Cambridge, MA: MIT Press.

Houk, J. C., & Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex, 2,* 95–110.

Hull, C. L. (1932). The goal gradient hypothesis and maze learning. *Psychological Review, 39,* 25–43.

Hull, C. L. (1934). The concept of the habit-family hierarchy and maze learning: Part I. *Psychological Review, 41,* 33–54.

Kelly, R. M., & Strick, P. L. (2004). Macro-architecture of basal ganglia loops with the cerebral cortex: Use of rabies virus to reveal multisynaptic circuits. *Progress in Brain Research, 143,* 449–459.

Killeen, P. R. (1994). Mathematical principles of reinforcement. *Behavioral and Brain Sciences, 17,* 105–172.

Klahr, D., Langley, P., & Neches, R. (Eds.). (1987). *Production system models of learning and development.* Cambridge, MA: MIT Press.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996, September 6). A neostriatal habit learning system in humans. *Science, 273,* 1399–1402.

Lewis, M. W., & Anderson, J. R. (1985). Discrimination of operator schemata in problem solving: Learning from examples. *Cognitive Psychology, 17,* 26–65.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology, 67,* 145–163.

Loewenstein, G., & Prelec, D. (1991). Negative time preference. *American Economic Review, 81,* 347–352.

Lovett, M. C. (1998). Choice. In J. R. Anderson & C. Lebiere (Eds.), *Atomic components of thought* (pp. 255–296). Hillsdale, NJ: Erlbaum.

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological Review, 104,* 241–265.

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

Mazur, J. E. (1985). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior: Vol. 5. The effect of delay and of intervening events on reinforcement value* (pp. 55–73). Hillsdale, NJ: Erlbaum.

Mazur, J. E. (2001). Hyperbolic value addition and general models of animal choice. *Psychological Review, 108,* 96–112.

Mazur, J. E. (2002). Concurrent-chain performance in transition: Effects of terminal-link duration and individual reinforcers. *Animal Learning Behavior, 30,* 249–260.

Mellers, B. A., Chang, S., Birnbaum, M. H., & Ordonez, L. D. (1992). Preferences, prices, and ratings in risky decision-making. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 347–361.

Meyer, D. E., & Kieras, D. E. (1997a). A computational theory of executive control processes and human multiple-task performance: Part 1. Basic mechanisms. *Psychological Review, 104,* 3–65.

Meyer, D. E., & Kieras, D. E. (1997b). A computational theory of executive control processes and human multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review, 104,* 749–791.

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictedness for reward responses in primate dopamine neurons. *Journal of Neurophysiology, 72,* 1024–1027.

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., Miotto, E. C., Feigenbaum, J. D., Bullock, P., & Polkey, C. E. (1997). Planning ability after frontal and temporal lobe lesions in humans: The effects of selection equivocation and working memory load. *Cognitive Neuropsychology, 14,* 1007–1028.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Myers, J. L., & Suydam, M. M. (1964). Gain, cost, and event probability as determiners of choice behavior. *Psychonomic Science, 1,* 39–40.

Myers, J. L., Suydam, M. M., & Gambino, B. (1965). Contingent gains and losses in a risk-taking situation. *Journal of Mathematical Psychology, 2,* 363–370.

Nason, S., & Laird, J. E. (2004). Soar–RL: Integrating reinforcement learning with Soar. In *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 208–213). Mahwah, NJ: Erlbaum.

Newell, A. (1973). Production systems: Models of control structures. In W. G. Chase (Ed.), *Visual information processing* (pp. 463–526). New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Niv, Y., Duff, M., & Dayan, P. (2005). Dopamine, uncertainty, and TD learning. *Behavioral and Brain Functions, 1,* 6.

O'Keefe, J. A., & Nadel, L. (1978). *The hippocampus as a cognitive map.* London: Oxford University Press.

O'Reilly, R. C. (2003). *Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia* (Tech. Rep. No. 03–03). Boulder: Institute of Cognitive Science, University of Colorado.

Packard, M. G. (1999). Glutamate infused posttraining into the hippocampus or caudate-putamen differentially strengthens place and response learning. *Proceedings of the National Academy of Sciences, USA, 96,* 12881–12886.

Packard, M. G., & McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory, 65,* 65–72.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience, 5,* 97–98.

Pirolli, P., & Fu, W.-T. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. In *Proceedings of the Ninth International Conference on User Modeling* (pp. 45–54). New York: Springer-Verlag.

Poldrack, R., Clark, J., Pare-Blagoev, E., Shohamy, D., Moyano, J., Myers, C., & Gluck, M. (2001, November 29). Interactive memory systems in the human brain. *Nature, 414,* 546–550.

Poldrack, R. A., Prabhakaran, V., Seger, C., & Gabrieli, J. D. E. (1999). Striatal activation during cognitive skill learning. *Neuropsychology, 13,* 564–574.

Rachlin, H. (2000). *The science of self-control.* Cambridge, MA: Harvard University Press.

Reid, A. K., & Staddon, J. E. R. (1998). A dynamic route finder for the cognitive map. *Psychological Review, 105,* 585–601.

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

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience, 13,* 900–913.

Schultz, W., Dayan, P., & Montague, P. R. (1997, March 14). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.

Schultz, W., Romo, R., Ljungberg, T., Mirenowica, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related signals carried by dopamine neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 233–248). Cambridge, MA: MIT Press.

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

Spence, K. W. (1932). The order of eliminating blinds in maze learning by the rat. *Journal of Comparative Psychology, 14,* 9–27.

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

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

Tadepalli, P., & Ok, D. (1998). Model-based average reward reinforcement learning. *Artificial Intelligence, 100,* 177–224.

Tolman, E. C., & Honzik, C. H. (1930). Degrees of hunger, reward and non-reward, and maze learning in rats. *University of California Publications in Psychology, 4*(16), 241–256.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76,* 31–48.

Tversky, A., & Russo, J. E. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical Psychology, 6,* 1–11.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14,* 101–118.

Watkins, C. J. (1989). *Learning from delayed rewards.* Unpublished doctoral thesis, University of Cambridge, England.

Williams, B. A. (1988). Reinforcement, choice, and response strength. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Steven's handbook of experimental psychology* (2nd ed., pp. 167–244). New York: Wiley.

Yeung, N., Botvinick, M., & Cohen, J. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111,* 931–959.

Yeung, N., & Sanfey, A. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience, 24,* 6258–6264.

(*Appendix follows*)

Appendix

Instructions Given to the Participants in the Maze-Searching Task

Please study the following instructions carefully before you begin.

This is a game in which you will be asked to make a series of choices, and you will be paid according to how many correct choices you can make. When the task begins, you will be situated in a room with a single object and two elements (an element is either "gold," "wood," "water," or "fire"). You have to choose one of the elements by clicking on its button. The object in the room is uniquely related to the correct element throughout the whole experiment.

After you select an element, you will be taken to another room. The room will be identical to the first room, except that the object and two elements are different. Again, you need to select an element by clicking on its button, which will take you to another room, and so on. When you have made three choices, a window will pop up to inform you whether the three choices you made are all correct or not. If all three choices are correct, you will be informed that you have successfully reached the end of the trial, and you will receive 3 points; another trial will then begin. On the other hand, if one of the three choices you made is wrong, you will be informed that you have reached a dead end, and 1 point will be deducted from your total score. Note that in this case you will only be informed that you have reached a dead end, but you will NOT be told which choice(s) you made is/are correct or wrong.

If you did not succeed, you may have made 1, 2, or 3 wrong choices (you will not be told how many wrong choices you have made). Before you can finish the trial you will need to get a total of 3 correct choices. When you reach a dead end, you need to click the "reset" button so that you can get a chance to make the needed number of correct choices. When you reset, you will be taken to another room and you can make another choice. You will then continue to make choices until (1) you make a total of three correct choices, in that case your trial will finish; or (2) you reach another dead end again, in that case you need to click "reset" again. However, you only need to make a total of three correct choices to finish a trial. For example, if in the first round you make one correct choice and two wrong choices and reach a dead end, after resetting you only need to make two more correct choices to reach the end (or if this time you make one or two wrong choices you will reach a dead end again).

There will be a total of 200 trials. In all the trials, all relations between the object in the room and the correct element will stay the same. Remembering the relations is therefore useful for making correct choices. You will be paid 1 cent per point in your total score in addition to the base payment of $8. The maximum you can earn is $15.

If you have any questions, please ask the experimenter NOW.

Received August 29, 2005
Revision received December 6, 2005
Accepted December 12, 2005 ■