Carnegie Mellon University
Research Showcase @ CMU

Department of Psychology | Dietrich College of Humanities and Social Sciences

November 2008

Extending the Computational Abilities of the Procedural Learning Mechanism in ACT-R

Wai-Tat Fu
Carnegie Mellon University

John R. Anderson
Carnegie Mellon University

Follow this and additional works at: http://repository.cmu.edu/psychology

This Conference Proceeding is brought to you for free and open access by the Dietrich College of Humanities and Social Sciences at Research Showcase @ CMU. It has been accepted for inclusion in Department of Psychology by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Extending the Computational Abilities of the Procedural Learning Mechanism in ACT-R

Wai-Tat Fu (wfu@cmu.edu)
John R. Anderson (ja+@cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

The existing procedural learning mechanism in ACT-R (Anderson & Lebiere, 1998) has been successful in explaining a wide range of adaptive choice behavior. However, the existing mechanism is inherently limited to learning from binary feedback (i.e. whether a reward is received or not). It is thus difficult to capture choice behavior that is sensitive to both the probabilities of receiving a reward and the reward magnitudes. By modifying the temporal difference learning algorithm (Sutton & Barto, 1998), a new procedural learning mechanism is implemented that generalizes and extends the computational abilities of the current mechanism. Models using the new mechanism were fit to three sets of human data collected from experiments of probability learning and decision making tasks. The new procedural learning mechanism fit the data at least as well as the existing mechanism, and is able to fit data that are problematic for the existing mechanism. This paper also shows how the principle of reinforcement learning can be implemented in a production system like ACT-R.

Introduction

Human choice behavior is often studied under various probability learning situations. In a typical probability learning situation, participants are asked to select one of the many options available, and feedback on whether the choice is correct or not is given after the selection. There are usually two main manipulations in a probability learning task: (1) the probabilities for each of the options being correct, and (2) the magnitudes of reward (usually monetary) received when the correct option is selected. One robust result is that people tend to choose the options a proportion of time equal to their probabilities of being correct – a phenomenon often called "probability matching" (e.g. Friedman et al., 1964). However, when the reward magnitudes are varied, the observed choice probabilities are sometimes larger or smaller than the outcome probabilities (e.g. Myers, Fort, Katz, & Suydam, 1963). These studies show consistently that people are sensitive to both outcome probabilities and reward magnitudes in making choices.

One limitation of the current ACT-R procedural learning mechanism (Lovett, 1998) is that it requires a pre-specification of correct and incorrect responses. Besides, feedback received is limited to a binary function (i.e. whether a reward is received or not). Apparently, a simple binary function may not be sufficient to represent the feedback from the environment. For example, imagine a situation in which there are several possible treatments for a particular disease and a physician has to choose a treatment that has the highest expected effectiveness. One may have to evaluate the effectiveness of each treatment through case-by-case feedback. For example, consider the case where the probabilities of effectiveness of three treatments 1, 2, and 3 are as shown in Figure 1. Since the effectiveness of each treatment follows a continuous distribution, a simple binary feedback function is obviously insufficient to represent the information received from the feedback.

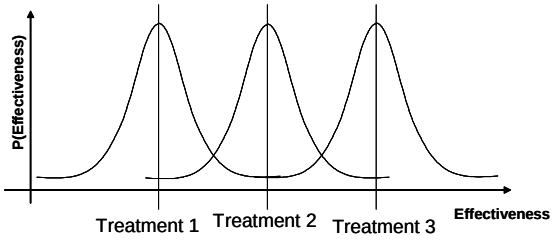

Figure 1. Probability of effectiveness of three treatments.

Another motivation for extending the current mechanism comes from recent findings of the functional role of dopaminergic signals in basal ganglia during procedural learning. Research shows that learning is driven by the deviation between the expected and actual reward (Schultz et al., 1995; Schultz, Dayan, & Montague, 1997). In other words, the reward magnitude is often processed as a scalar quantity – depending on whether the magnitude of the actual reward is higher or lower than expected, a positive or negative reinforcement signal is generated respectively. The pre-specification of correct and incorrect responses is therefore inconsistent with the current understanding of the procedural learning mechanism in basal ganglia.

The ACT-R 5.0 architecture

Figure 2 shows the basic architecture of the ACT-R 5.0 system. The core of the system is a set of production rules that represents procedural memory. Production rules coordinate actions in each of the separate modules. The modules communicate to each other through its buffer, which holds information necessary for the interaction between the system and the external world. Anderson, Qin, Sohn, Stenger, and Carter (2003) showed that the activity in these buffers match well to the activities in certain cortical areas (see Figure 2). The basal ganglia are hypothesized to implement production rules in ACT-R, which match and act on patterns of activity in the buffers. This is consistent with a typical ACT-R cycle in which production rules are matched to the pattern of activity in the buffers, a production is selected and fired, and the contents in the buffers updated. In ACT-R, when there is more than one production matching the pattern of buffer activity, the system selects a production based on a conflict resolution mechanism. The basis of the conflict resolution mechanism is the computation of expected utility, which captures the

effectiveness and efficiency of the production in accomplishing the goals of the system.

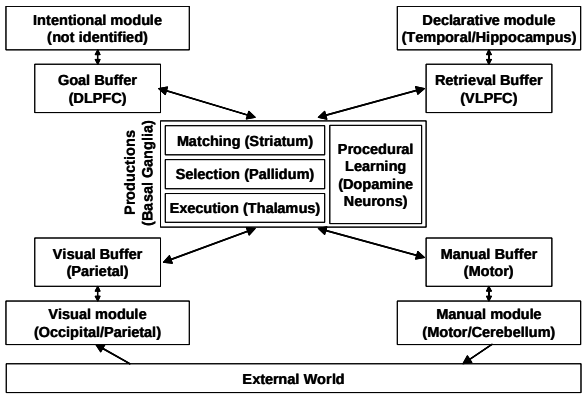

Figure 2. The ACT-R 5.0 architecture.

To adapt to the environment, the system must learn from the consequences of its actions so that when the same conditions are met in the future, a better choice of productions can be made. Procedural learning updates the expected utility of a production from the consequences of firing the production, and the dopamine systems in basal ganglia are believed to be involved in the learning process. Specifically, procedural learning appears to be coded by dopaminergic signals from the ventral tegmental area (VTA) and substatntia nigra to the striatum in basal ganglia (Schultz, et al., 1995; Schultz, et al., 1997), and different patterns are either reinforced or penalized according to the dopaminergic signals. Previous studies (Ljungberg, Apicella, & Schultz, 1992; Mirenowicz, Schultz, 1994) show that the activation of dopamine neurons depends entirely on the difference between the predicted and actual rewards. Once an unpredicted reward is perceived, response in dopamine neurons is transferred to the earlier reward-predicting stimulus. Inversely, when a predicted reward fails to occur, dopamine neurons are depressed in their activity at exactly the time when the reward would have occurred (Schultz, Apicella, Ljungberg, 1993). It therefore appears that dopamine signals do not simply report the occurrence of rewards. Rather, outputs from dopamine neurons appear to code for a deviation or error between the actual reward received and predictions or expectations of the reward. In other words, dopamine neurons seem to be feature detectors of the "goodness" of environmental events relative to the learned expectations about those events.

The current procedural learning mechanism

During each cycle of ACT-R, productions that match the contents of the buffer will be put into a conflict set. The productions in the conflict set are ordered in terms of their expected utility and ACT-R considers them according to that ordering. The expected utility of a production is defined as E = PG-C, where P is the estimated probability that the goal will be achieved if that production is chosen, G is the value of the goal, and C is the estimated cost of achieving the goal if that production is chosen (see Table 1). Procedural learning updates the value of P and C according to the following equations:

$$P = \frac{successes}{successes + failures} \qquad C = \frac{efforts}{successes + failures},$$ where

successes and *failures* are the number of times the production has succeeded or failed to accomplish the current goal respectively (i.e. a reward or penalty), and *efforts* is the total amount of time taken over all past uses of the production rule, successful or failed. These quantities start out with initial values that are updated with experience. For example, if for production n the initial *successes* equals 1, *failures* equals 1, and *efforts* equals 0.5, when a pre-specified *success* is encountered 0.1 second after k has fired, P will change from 0.5 to 0.67 (=2/(2+1)), C will change from 0.25 to 0.2 (=(0.5+0.1)/(2+1)). If G equals 20, then the expected utility (E=PG-C) will increase from 9.75 (=0.5*20-0.25) to 13.13 (=0.67*20-0.2). The successful experience has thus acted as a reward and reinforced production n by increasing its expected utility, and as a consequence, n will be more likely to be selected in the future..

Table 1. A list of free parameters and their definitions.

Parameters	Definition (Old mechanism)
G	Value of the goal (measured in seconds)
successes/ failures	Initial number of times the production has led to a success/failure state before the model starts
efforts	Total amount of time taken over all past uses of the production, successful or failed.
Parameters	**Definition (New mechanism)**
r_n	The actual reward received
K	The discount factor ($0 < K \leq 1$). Future rewards are discounted by 1/(1+KD), where D is the time between the firing of the current and the next production.
a	The learning rate.
D_{n+1}	The time between the consecutive firing of production n and n+1

Although the existing mechanism was able to match to human choice behavior, there are aspects in which the mechanism can be improved. First, in the existing mechanism, learning of P requires pre-specification of successful or failure states and the expected utility will increase or decrease respectively when the state is reached. The use of success and failure states may not be sufficient in situations where a continuous feedback function is required. From a practical perspective, pre-specification of success and failure states could be difficult especially in complex tasks, in which some states are often considered "more successful" than others. One way to improve the current mechanism is to learn from a scalar reward value. Being able to assign a scalar reward value to a production therefore allows more flexible pre-specification of the reward structure of the environment and allows the model to adapt to the environment accordingly. Second, the existing

procedural learning mechanism will change the expected utilities of productions only when the actual outcome is experienced, which requires keeping track of the whole sequence of previous productions that leads to the outcome. This could be computationally expensive especially when the number of productions is large. It is therefore desirable to have a mechanism that learns from local information before the outcomes are known.

The new procedural learning mechanism

In the artificial intelligence community, algorithms have been developed to allow agents to learn in different environments (Sutton & Barto, 1998). One established algorithm is the Temporal Difference (TD) algorithm, which was originally inspired by behavioral data on how animals learn prediction (Sutton & Barto, 1981). Research showed that the TD algorithm is well suited to explain the functional role of dopaminergic signals (e.g. Houk, et al., 1995; Holroyd & Coles, 2002, O'Reilly, 2003). The TD algorithm is designed to learn to estimate future rewards based on experience, and has a built-in credit assignment mechanism that reinforces the predicting stimuli.

In its simplest form, the new mechanism can be represented as $U'(n) = U(n) + aTD(n)$, where $U'(n)$ is the updated value of the expected utility $U(n)$ of production n after an ACT-R cycle, a is the learning rate, and $TD(n)$ is the temporal difference error. $TD(n)$ calculates the difference between the actual and expected rewards, i.e. $TD(n) = R(n) - U(n)$. The basic learning mechanism is therefore similar to the learning rule of Rescola and Wagner (1972) (e.g. see Sutton & Barto, 1981). The measure of future rewards has to take into account long-term as well as short-term consequences. It is plausible to weigh immediate primary reinforcement more strongly than delayed primary reinforcement. We chose to use the hyperbolic function to discount delayed reinforcement (the justification of using the hyperbolic function is beyond the scope of this paper, but see Lowenstein & Prelec, 1991; Mazur, 2001). A good estimate of the total future rewards is therefore $R(n) \approx r_n + U(n+1)/(1+KD_{n+1})$, where r_n is the immediate reward received for production n, $U(n+1)$ is the expected utility of the production that fires after production n, K is the discount parameter, and D_{n+1} is the time lag between the times when production n and production n+1 fire. To implement the mechanism in ACT-R, the basic algorithm has to be modified to take both the reward and cost into account and translate them into a single dimension[1] – i.e. the reinforcement will be the difference between the reward and cost (i.e. the net reward). In other words, the estimate becomes $R(n) \approx r_n - C_n + U(n+1)/(1+KD_{n+1})$, where C_n is the cost of firing production n. Putting the estimate of R(n) back to the equation for $U'(n)$, we have:

$$U'(n) = U(n) + a[r_n - C_n + U(n+1)/(1+KD_{n+1}) - U(n)]$$

One can see that when the estimate is perfectly accurate, $TD(n) = 0$, or $U(n) = r_n - C_n + U(n+1)/(1+KD_{n+1})$ and learning will stop. The value of TD(n) can therefore be considered the prediction error (as encoded by dopaminergic signals), and the mechanism learns by reducing this prediction error. It can easily be seen that once a primary reward is received, the expected utility of the productions that lead to the reward will be credited with a discounted reward, and discounting is heavier the farther away the production is from the reward.

The new mechanism updates the expected utility based on the difference between the predicted and actual net reward. There are two main differences between the new and existing mechanisms. In the new mechanism, the reward is a scalar quantity, and the amount of change is determined by the difference between the predicted and actual reward, which is consistent with the functional role of dopaminergic signals. This characteristic allows the new mechanism to extend its learning capabilities beyond a binary function as in the existing mechanism. Second, in the existing mechanism, learning requires keeping track of a long sequence of productions that lead to the reward. However, in the new mechanism, only the expected utility of the next production is required. The reinforcement signal will eventually propagate back to the productions that lead to the reward.

Testing the new mechanism

The goal of this paper is to show the limitations of the existing mechanism and how the new mechanism is able to extend the learning capabilities of ACT-R. However, owing to space limitation, we are unable to show all properties of mechanism. For example, none of the data sets in this paper was sensitive to the discount parameter K, so we fixed it at 1.0 and just varied the value of r_n to fit the data[2]. The learning rate a was also fixed at 0.1. We first used the new mechanism to fit two data sets from the probability learning tasks by Friedman et al. (1964) and Myers et al. (1963). Since these two sets of data were also modeled well by the existing mechanism (Lovett, 1998), we were able to compare the results of the two mechanisms and show that the use of TD error to drive the learning process is at least as effective as the existing mechanism. Finally, we used the new mechanism to fit the data from a decision making task studied by Busemeyer and Myung (1992), which we believe were problematic for the existing mechanism.

Probability matching behavior

In Friedman et al., participants completed more than 1,000 choice trials over the course of three days. For each trial, a signal light was illuminated, participants pressed one of the two buttons, and then one of the two outcome lights was

[1] ACT-R takes the agnostic economist's position of simply assuming these map onto some internal values without deeply inquiring why.

[2] Since the delay D is a constant for all data sets, it can be shown that the parameter K is absorbed into the value or r_n.

illuminated. Task instructions encouraged participants to try to guess the correct outcome for each trial. The study extended the standard probability learning paradigm by changing the two buttons' success probabilities across 48-trial blocks during the experiment. Specifically, for the odd-numbered blocks 1-17, the probabilities of success of the buttons (p and 1-p) were 0.5. For the even-numbered blocks 2-16, p took on the values from 0.1, to 0.9 in a random order. We focus on the analysis of the even-numbered blocks, as they show how people adapted to the outcomes with experience.

Table 2. Observed and predicted choice proportions from the experiment by Friedman et al. (1964). Predicted scores are in parentheses. Each block has 12 trials.

	Probabilities			
P	Block 1	Block 2	Block 3	Block 4
0.1	0.34 (0.37)	0.23 (0.24)	0.18 (0.17)	0.15 (0.13)
0.2	0.37 (0.41)	0.26 (0.26)	0.29 (0.23)	0.31 (0.23)
0.3	0.49 (0.49)	0.41 (0.41)	0.44 (0.34)	0.35 (0.33)
0.4	0.46 (0.53)	0.44 (0.50)	0.38 (0.43)	0.38 (0.38)
0.6	0.56 (0.59)	0.51 (0.59)	0.52 (0.55)	0.52 (0.57)
0.7	0.50 (0.56)	0.53 (0.64)	0.58 (0.72)	0.62 (0.75)
0.8	0.50 (0.51)	0.76 (0.71)	0.74 (0.77)	0.73 (0.78)
0.9	0.66 (0.62)	0.78 (0.79)	0.78 (0.81)	0.79 (0.81)

Table 2 shows the observed and predicted proportion of choices in the experiment by Friedman et al. Participants in general exhibited probability matching behavior. Across the four 12-trial subblock, participants chose the correct buttons in roughly 50% of the trials in the first block and approached the corresponding p values in each block. The predicted proportions were generated by the model, which had two critical productions, *Choose-Right-Button* and *Choose-Left-Button*, and the expected utilities of these productions were learned according to the new mechanism. The exact sequence of outcomes as reported in Friedman et al. was presented to the model. A reward of 3 is obtained when the correct button was chosen (i.e. r_n=3). The initial expected utilities of the two productions were set to 0. The fit was good, R^2 = 0.97, MSE = 0.003, which was similar to the model based on existing procedural learning mechanism. We conclude that the new mechanism can represent the learning mechanism at least as well as the existing mechanism with the same number of free parameters.

Overmatching behavior

Myers et al. performed another probability learning experiment, but they also varied the amount of monetary reward that participants would receive for each correct response. Participants would either receive no reward or penalty, ±1¢, or ±10¢ for each correct and incorrect responses. The probabilities that one of the alternatives was correct were p=0.6, p=0.7, and p=0.8. Table 3 shows the choice proportions for the participants in each of the conditions. When there was no reward, participants seemed to be exhibiting probability matching behavior. However, when there was a monetary reward, participants seemed to be "overmatching". From the data, it also appears that the higher the reward, the more the choice proportion exceeds the matching probability.

Table 3. Observed and predicted choice proportions from the experiment by Myers et al. (1963). Predicted scores are in parentheses.

Reward	Probabilities		
(cents)	p = 0.6	p = 0.7	p = 0.8
0	0.624 (0.612)	0.753 (0.750)	0.869 (0.829)
1	0.653 (0.676)	0.871 (0.834)	0.925 (0.938)
10	0.714 (0.711)	0.866 (0.836)	0.951 (0.944)

Since the task is basically the same as in Friedman et al., we used the same model to fit the data. We used the same set of parameters to fit the data in the no reward conditions (i.e. reward = 3). We chose the reward parameters (reward=±4.97 and ±5.7 for the ±1¢ and ±10¢ conditions respectively[3]) in the reward conditions to maximize the fit, and obtained R^2 of 0.98 and MSE of 0.0008, which is similar to the fit obtained by the existing procedural learning mechanism. However, we had only two free parameters in this model, compared to three free parameters in the model reported in Lovett (1998). In addition, the new mechanism provides a more natural interpretation of the overmatching behavior – when the reward was large, learning increases the expected utilities of the successful productions to higher values (since the deviation was larger). As a consequence, the model exhibited overmatching behavior. On the other hand, Lovett (1998) manipulated the architectural parameter G to fit the data, which seems awkward, as G is not supposed to be directly under strategic control.

Learning from normally distributed rewards

Busemeyer and Myung (1992) conducted 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, and the next less effective treatment is called Treatment 2, and so on (see Figure 1). 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 based on trial-by-trial feedback. Participants were told to maximize the sum of the treatment effects over training and they were paid 4¢ per point. The means of the normal distributions are m-d, m and m+d for Treatment 1, 2, and 3 respectively. The two independent variables were mean difference (d) (i.e. the

[3] Since the reward values used in the model reflect subjective values, they do not necessarily follow a linear relationship with the external reward values.

separation of the distributions in Figure 1) and standard deviation (s) (which affects the amount of overlap in Figure 1). The exact values of d and s are shown in Table 4. Each participant was given 9 blocks (50 trials per block) of training in each condition. The model received the same amount of training as the participants.

From Table 4, we can see that as the mean difference increased, the observed choice proportions of the optimal treatment increased. As the standard deviation increased, the observed choice proportions of the best treatment decreased except when the mean difference was 3.0. The results showed that participants adapted their choice by learning the expected effectiveness of treatments. The results also showed that the more distinguishable the distributions were (larger mean difference or smaller standard deviation), the more likely the participants would choose the best treatment.

Table 4. Observed and predicted choice proportions of the optimal treatment from the experiment by Busemeyer & Myung (1992). Predicted scores are in parentheses.

Standard deviation (s)	Mean difference (d)		
	2.0	2.5	3.0
3.0	0.69 (0.74)	0.84 (0.79)	0.85 (0.84)
4.5	0.69 (0.72)	0.72 (0.76)	0.84 (0.80)
6.0	0.65 (0.68)	0.63 (0.69)	0.86 (0.83)

To model the data, we built three productions that chose each of the treatments. The initial expected utility 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. The value of r was chosen to be 1.76 to best fit the data. We obtained a fit of R^2=0.94, RMSE=0.007. The good fit to the data show that the new learning mechanism was able to build up the expected effectiveness of the treatments from trial-by-trial feedback, and was able to exhibit similar sensitivity to the differences of the distributions as participants. Since the effectiveness was sampled from a normal distribution, it is difficult to pre-specify which treatment was successful. It is therefore difficult to use the existing learning mechanism to model these data. In the new mechanism, however, whenever the actual reward was higher than the expected utility of the production, the production will be reinforced; otherwise the production will be penalized. With the same amount of experience (50 trials), the expected utilities of the production were able to reflect the actual expected effectiveness of the treatments.

Summary

We have fit a new procedural learning mechanism of ACT-R to three separate sets of data with all parameters held constant except the reward magnitudes the models received after each trial. In the first two cases, the new mechanism did at least as well as the existing mechanism in capturing the observed choice proportions in different settings. In the last case, we showed that the new mechanism fits data that are problematic for the existing mechanism. The new mechanism learned to probability match the true probabilities of outcomes by reducing the difference between the expected and actual reward. As the difference diminished, the change in the prediction decreased. When the reward was large, learning increases the expected utilities of the successful productions to higher values (since the deviation was larger). As a consequence, the chance of selecting the option that had the higher probability of being correct increased – i.e. the model exhibited overmatching behavior.

Although the first two sets of data can be modeled by the existing learning mechanism, the new mechanism provided a more natural explanation to the results. In the final set of data, we showed how the new mechanism generalizes and extends the computational abilities of the existing mechanism. The mechanism was able to learn the expected effectiveness of each treatment based on trial-by-trial feedback, without the need to pre-specify whether the productions had led to successful or failure states.

Discussion

We have presented a new procedural learning mechanism in ACT-R. The use of the deviation between the expected and actual reward values in the new learning mechanism is consistent with the current understanding of the functional role of VTA dopamine neurons in basal ganglia. We showed that the new mechanism generalizes and extends the computational abilities of the existing procedural learning mechanism. Specifically, the new mechanism is not limited to learning from binary feedback functions. Rather, the new mechanism is able learn from continuous reward functions with similar sensitivity to the variations in the reward distributions. The current paper also showed how the reinforcement learning mechanism observed in basal ganglia can be implemented in production systems such as ACT-R.

In practice, the current mechanism allows the use of a scalar reward parameter without the need to pre-specify success or failure states in a task. This pre-specification could be difficult especially in complex tasks in which a state could sometimes be good or bad depending on one's experience with the task, as experience may change one's expectation of different states. In addition, although the existing mechanism can adapt to different magnitudes of reward, the change of the architectural parameter G (in E=PG-C) to fit the data may not be easy in complex tasks that has many subgoals, especially when some subgoals may be considered "more successful" than the others.

Owing to space limitations, we are not able to show all properties of the mechanism. In fact, we have only tested the mechanism in single-choice tasks, which did not depend critically on the credit assignment mechanism. The discounting of future rewards therefore did not affect performance of the models in all three tasks that we have presented. However, we believe the discounting mechanism

is more plausible than the existing mechanism, in which immediate and future rewards are weighted equally.

In all three data sets, the model had the same amount of experiences as the participants and reached the same level of asymptotic performance. In the first data set, we also showed that the performance of the model in each of the four subblocks matched the participants well, suggesting that the learning rate of the mechanism is comparable to that of the participants. However, it is possible that the reinforcement learning mechanism could be slow for more complex tasks. It could be problematic, for example, when a primary reward is received after a long sequence production firings. Since only one production is updated during each ACT-R cycle, the primary reward may take several cycles to propagate back to the production where the critical decision is made. It is not clear how people learn in such situations. It is possible that they rely on direct instruction to point out such contingencies rather than counting on a automatic learning mechanism. It does not seem that the mechanisms behind the dopamine reward system are capable of spanning unbounded lengths of time in a way that would lead to rapid convergence.

Acknowledgments

The current work is supported by a grant from the office of naval research (N00014-99-1-0097). We thank Niels Taatgen, Pat Langley, and two anonymous reviewers for the useful comments on an earlier version of this paper.

References

Anderson, J. R. (1990). *Rules of the mind.* Mahwah, NJ:Erlbaum.

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

Anderson, J. R., Qin, Y., Sohn, M.-H., Stenger, V. A., Carter, C. S. (2003). An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic Bulletin & Review, 10* (2), 241-261.

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* (2), 177-194.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*, 747-749.

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.

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* (4), 679-709.

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.

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. *The American Economic Review, 81* (2), 347-352.

Lovett, M. C. (1998). Choice. In C. Lebiere and J. R. Anderson, *Atomic components of thought* (chapter 8, pp. 255-296).

Mazur, J. E. (2001). Hyperbolic value addition and general models of animal choice. *Psychological review, 108* (1), 96-112.

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

Myers, J. L., Fort, J. G., Katz, L., & Suydam, M. M. (1963). Differential monetary gains and losses and event probability in a two-choice situation. *Journal of Experimental Psychology, 66*, 521-522.

O'Reilly, R. C. (2003). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Institute of Cognitive Science, University of Colorado, Boulder, Technical Report 03-03.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: variation in the effectiveness of reinforcement and non reinforcement. In Black, A. H. and Prokasy, W. F. (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). A neural substrate of prediction and reward., *Science, 275*, 1593.

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.

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

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