The Journal of Neuroscience, March 1, 1996, 16(5):1936–1947

A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning

P. Read Montague,[1] Peter Dayan,[2] and Terrence J. Sejnowski[3,4]

[1]Division of Neuroscience, Baylor College of Medicine, Houston, Texas 77030, [2]CBCL, Department of Brain and Cognitive Science, Cambridge, Massachusetts 02139, [3]The Howard Hughes Medical Institute and The Salk Institute for Biological Studies, La Jolla, California 92037, and [4]The Department of Biology, University of California at San Diego, La Jolla, California 92093

We develop a theoretical framework that shows how mesencephalic dopamine systems could distribute to their targets a signal that represents information about future expectations. In particular, we show how activity in the cerebral cortex can make predictions about future receipt of reward and how fluctuations in the activity levels of neurons in diffuse dopamine systems above and below baseline levels would represent *errors* in these predictions that are delivered to cortical and subcortical targets. We present a model for how such errors could be constructed in a real brain that is consistent with physiological results for a subset of dopaminergic neurons located in the ventral tegmental area and surrounding dopaminergic neurons. The theory also makes testable predictions about human choice behavior on a simple decision-making task. Furthermore, we show that, through a simple influence on synaptic plasticity, fluctuations in dopamine release can act to change the predictions in an appropriate manner.

Key words: prediction; dopamine; diffuse ascending systems; synaptic plasticity; reinforcement learning; reward

In mammals, mesencephalic dopamine neurons participate in a number of important cognitive and physiological functions including motivational processes (Wise, 1982; Fibiger and Phillips, 1986; Koob and Bloom, 1988), reward processing (Wise, 1982), working memory (Sawaguchi and Goldman-Rakic, 1991), and conditioned behavior (Schultz, 1992). It is also well known that extreme motor deficits correlate with the loss of midbrain dopamine neurons; however, activity in the substantia nigra and surrounding dopamine nuclei, i.e., areas A8, A9, A10, does not show any systematic relationship with the metrics of various kinds of movements (Delong et al., 1983; Freeman and Bunney, 1987).

Physiological recordings from alert monkeys have shown that midbrain dopamine neurons respond to food and fluid rewards, novel stimuli, conditioned stimuli, and stimuli eliciting behavioral reaction, e.g., eye or arm movements to a target (Romo and Schultz, 1990; Schultz and Romo, 1990; Ljungberg et al., 1992; Schultz, 1992; Schultz et al., 1993). Among a number of findings, these workers have shown that transient responses in these dopamine neurons transfer among significant stimuli during learning. For example, in a naive monkey learning a behavioral task, a significant fraction of these dopamine neurons increase their firing rate to unexpected reward delivery (food or fluid). In these tasks, some sensory stimulus (e.g., a light or sound) is activated so that it consistently predicts the delivery of the reward. After the task has been learned, few cells respond to the delivery of reward and more cells respond to the onset of the stimulus, which is the predictive sensory cue (see Figs. 1, 2). More important, in these and similar tasks, activity levels in these neurons are sensitive to the precise time at which the reward is *delivered* after the onset of the predictive sensory cue.

The capacity of these dopamine neurons to represent such predictive temporal relationships and their well described role in reward processing suggest that mesolimbic and mesocortical dopamine projections may carry information related to expectations of future rewarding events. In this paper, we present a brief summary of the physiological data and a theory showing how dopamine neuron output could, in part, deliver information about predictions to their targets in two distinct contexts: (1) during learning, and (2) during ongoing behavioral choice. Under this theory, stimulus–stimulus learning and stimulus–reward learning become different aspects of the same general learning principle. The theory also suggests how information about future events can be represented in ways more subtle than tonic firing during delay periods.

First, we describe the physiological and behavioral data that require explanation. Second, we develop the theory, show its equivalence to other algorithms that have been used for optimal control, and demonstrate how and why it accounts for dopamine neuron output during learning tasks. Third, using the theory, we generate predictions of human choice behavior in a simple decision-making task involving a card choice experiment.

DOPAMINERGIC ACTIVITY

In a series of experiments in alert primates, Schultz and colleagues have shown how neurons in dopaminergic nuclei fire in response to sensory stimuli and the delivery of reward (Romo and Schultz, 1990; Schultz and Romo, 1990; Ljungberg et al., 1992; Schultz et al., 1993). These neurons provide dopaminergic input to widespread targets including various limbic structures and the prefron-

Received Aug. 21, 1995; revised Nov. 28, 1995; accepted Dec. 6, 1995.

This work was supported by NIMH Grant R01MH52797 and the Center for Theoretical Neuroscience at Baylor College of Medicine (P.R.M.), SERC (P.D.), and the Howard Hughes Medical Institute and NIMH Grant R01MH46482–01 (T.J.S.). We thank Drs. John Dani, Michael Friedlander, Geoffrey Goodhill, Jim Patrick, and Steven Quartz for helpful criticisms on earlier versions of this manuscript. We thank David Egelman for comments on this manuscript and access to experimental results from ongoing decision-making experiments.

Correspondence should be addressed to P. Read Montague, Division of Neuroscience, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030.

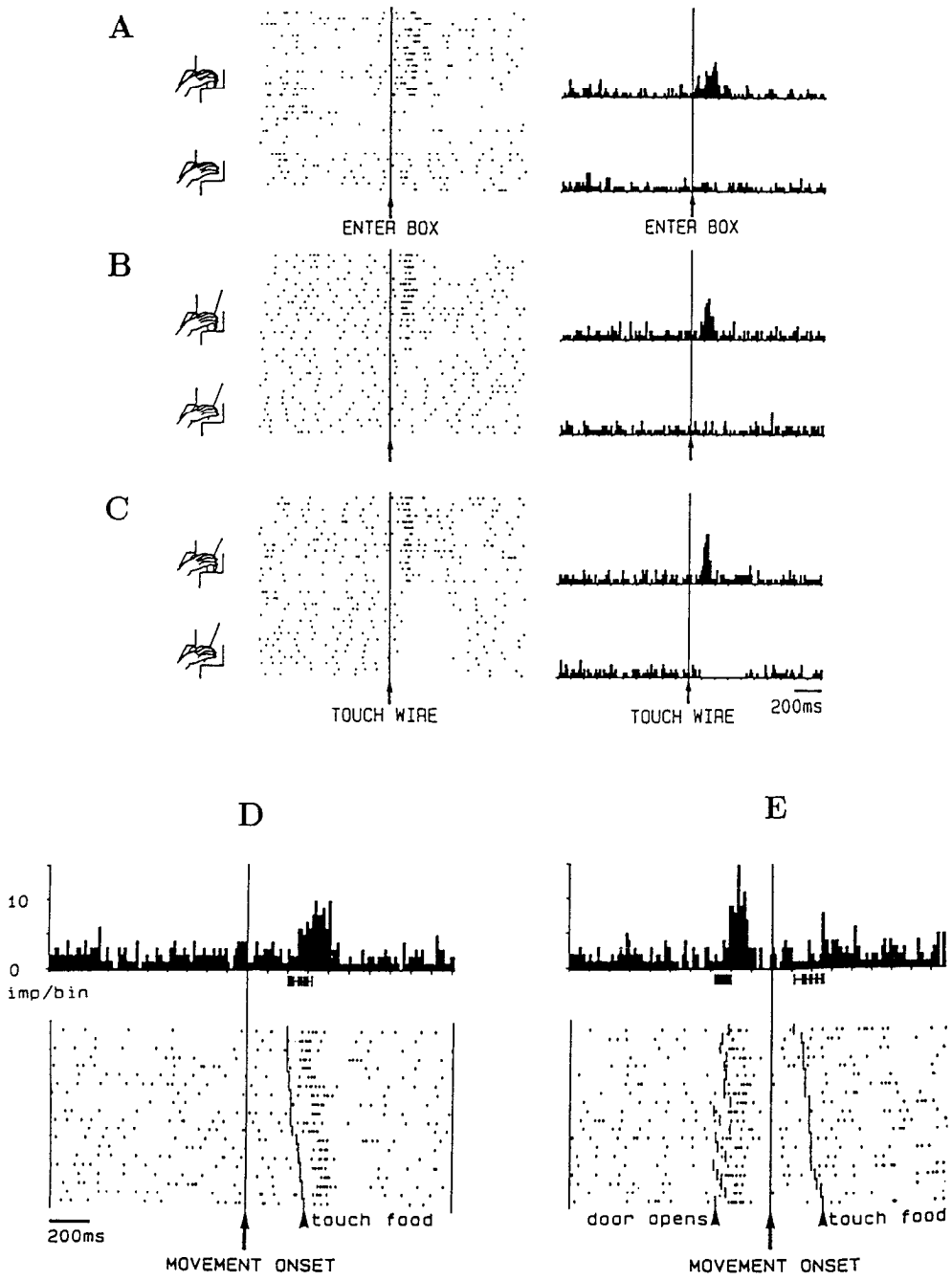

Figure 1. Object specific responses of dopamine neurons: self-initiated and triggered movements. Animal is trained to reach into a visually occluded food box in response to the sight or sight and sound of food box door opening. A food morsel was present in the box and was connected to a touch-sensitive wire in some test conditions. *A–C* show responses in 3 dopamine neurons for self-initiated movements: perievent histograms are shown at the *far right*, raster plots of individual trials in the *middle*, and illustration of test conditions on the *far left*. Transient increases in firing rate occurred only after touching food and not during arm movement, exploration of empty food box, or touching of bare wire. *A*, Food morsel touch versus search of empty box (trials aligned to entry into box). *B*, Same as *A* except food stuck to end of touch-sensitive wire versus only wire. *C*, Same as *B*. Depression in activity occurs after wire touch. *D*, Response of dopamine neuron to food touch. Movement self-initiated. *E*, Response of dopamine neuron to door opening with no response to food touch. Movement triggered by door opening. In *D* and *E*, the plots have been aligned to movement onset. (*A–E* reproduced with permission from Romo and Schultz, 1990.)

tal cortex (Oades and Halliday, 1987). One of these nuclei, the ventral tegmental area (VTA), and one of its afferent pathways, the medial forebrain ascending bundle, are also well known self-stimulation sites (Wise and Bozarth, 1984).

Object-specific dopamine neuron responses unrelated to movement parameters

Figure 1*A–C*, reproduced from Romo and Schultz (1990), shows the responses of mesencephalic dopamine neurons in two conditions: (1) self-initiated arm movements into a covered, food box without triggering stimuli, and (2) arm movements into the food box triggered by the rapid opening of the door of the box. In the latter condition, the door opening was either visible and audible or just audible. The animals were first trained on the trigger stimulus, i.e., while the animal rested its hand on a touch-sensitive lever, the food box door opened, the animal reached into the box, grabbed a piece of apple, and ate it. The piece of apple was stuck to the end of a touch-sensitive wire. After this task had been learned, the self-initiated movement task was undertaken. The recordings shown in Figure 1 are from three dopamine neurons contralateral to the arm used in the task. These and other control experiments from this paper show that under the conditions of this experiment (1) these dopamine neurons give a transient response if a food morsel is felt, (2) arm movement alone does not inhibit or activate the subsequent firing of the dopamine neurons, and (3) simply touching an object (the bare wire) is not sufficient to yield the transient increase in firing. Ipsilateral dopamine neurons yielded the same results. Over 80% of the neurons recorded showed this qualitative behavior.

The responses change completely if a stimulus consistently precedes (triggers) reaching into the food box. After learning,

when movement of the arm to the food box was triggered by a sensory stimulus (door opening as described above), 77% of the dopamine neurons gave a burst after door opening and gave no response to the touch of food in the box. This is shown in Figure 1,*D* and *E*, also reproduced from Romo and Schultz (1990).

In a series of related tasks with the same monkeys, Schultz and Romo (1990) showed that monkeys react to door opening with target directed saccades. The response of the dopamine neurons was specific to the multimodal sensory cues associated with door opening because dopamine neurons also responded to door opening during the absence of eye movements (eye already on target when door opened). Moreover, sensory cues associated with door opening did not cause dopamine neurons to fire outside the context of the behavioral task.

Dopamine neuron access to temporal information

Reaction-time task

In Ljungberg et al. (1992), a light would come on signaling that the monkey should move its hand from a resting key to a lever. A reward consisting of a mechanically delivered drop of fruit juice would be delivered 500 msec after pushing the lever. During early learning trials, there was little extra firing in the dopamine neurons after the light came on but, when juice was given to the monkey, dopamine cells transiently increased their firing (Ljungberg et al., 1992). After the animal had learned this reaction-time task, the onset of the light caused increased dopamine activity; however, the delivery of the juice no longer caused significant change in firing. Similar to the above results, the transient responses of the dopamine neurons transferred from reward delivery to light onset.

Spatial-choice tasks

Monkeys trained on the reaction-time task described above were subsequently given three tasks in which one of two levers was depressed to obtain a juice reward (spatial choice task in Schultz et al., 1993) (Fig. 2). Each lever was located underneath an instruction light that would indicate which lever to depress. The delivery of the reward followed a correct lever press by 500 msec so that dopamine neuron responses to lever touch could be distinguished from responses to reward delivery. Dopamine neuron responses for the three tasks are shown in Figure 2. As explained in the legend, the difference between separate tasks was the temporal consistency between the instruction and trigger lights. As with the reaction-time task (Ljungberg et al., 1992) and the triggered task above (Romo and Schultz, 1990; Schultz and Romo, 1990), dopamine neuron responses transferred from reward delivery to sensory cues that predicted reward.

These experiments (Schultz et al., 1993) show that the dopamine neurons have access to the expected time of reward delivery. Figure 3 shows the response of a single dopamine neuron during the delayed response task. The response of this single neuron is shown in the presence and absence of reward delivery. These results were obtained while the animal was still learning the task. When no reward was delivered for an incorrect response, only a depression in firing rate occurred at the time that the reward would have been delivered.

These results of Schultz and colleagues illustrate four important points about the output of midbrain dopamine neurons. (1) The activities of these neurons do not code simply for the time and magnitude of reward delivery. (2) Representations of both sensory stimuli (lights, tones) and rewarding stimuli (juice) have access to driving the output of dopamine neurons. (3) The drive

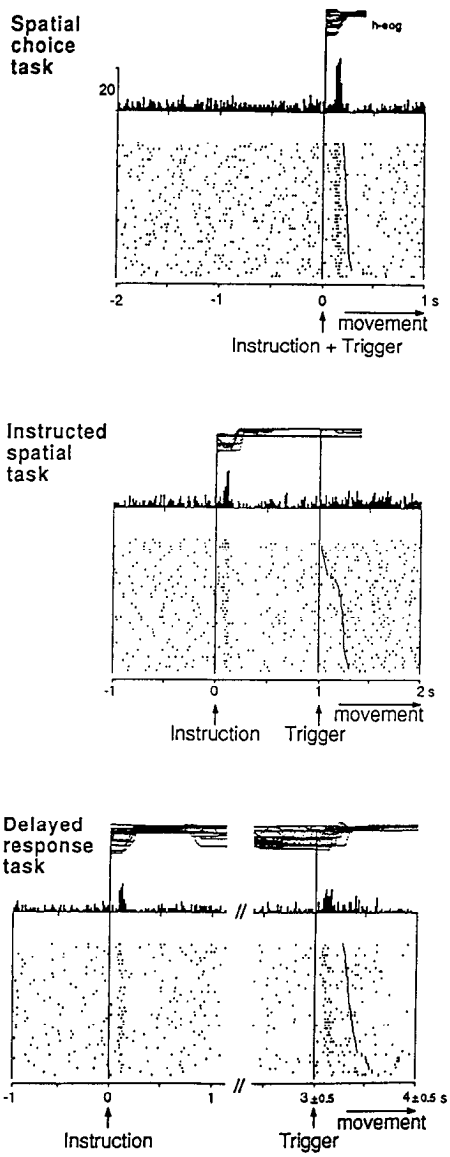

Figure 2. Spatial choice tasks (after learning). The animal sits with hand resting on a resting key and views two levers (medial and lateral) located underneath two green instruction lights. These lights indicate which lever is to be pressed once a centrally located trigger light is illuminated. Three separate tasks were learned. The main difference among the tasks was the temporal relationship of instruction light and trigger light illumination. The three tasks were called spatial choice task (*A*), instructed spatial task (*B*), and spatial delayed response task (*C*). This figure, reproduced from Schultz et al. (1993), shows the responses of 3 dopamine neurons during task performance (after training). *A*, Spatial choice task: the instruction and trigger lights were illuminated together; the animal released a resting key and pressed the lever indicated by the instruction light. *B*, Instructed spatial task: the instruction light came on and stayed on until the trigger light came on exactly 1 sec later. *C*, Spatial delayed response task: the instruction light came on for 1 sec and went out. This was followed by the illumination of the trigger light with a delay randomly varying between 1.5 and 3.5 sec (indicated by *broken lines*). In all tasks, lights were extinguished after lever touch or after 1 sec if no movement occurred. 0.5 sec after a correct lever press, reward (mechanically delivered juice) was delivered. The three panels show cumulative histograms with underlying raster plot of individual trials. The onset of arm movement is indicated by a *horizontal line*, and horizontal eye movements are indicated by *overlying traces*. Each panel shows data for 1 neuron. The vertical scale is 20 impulses/bin (indicated in *A*). Reproduced from Schultz et al. (1993) with permission from *The Journal of Neuroscience*.

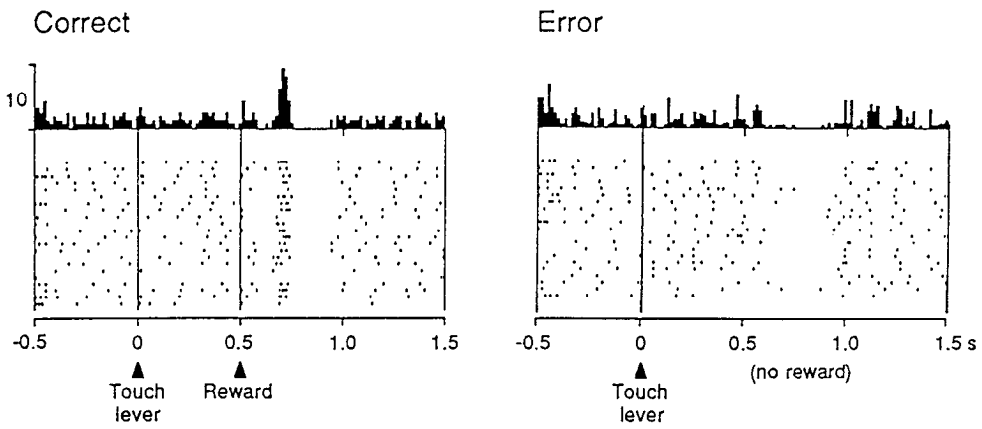

Figure 3. Timing information available at the level of dopamine neurons. Transient activation is replaced by depression of firing rate in a single dopamine neuron during error trials, i.e., animal depresses incorrect lever during acquisition of spatial delayed task. *Left*, Transient increase in firing rate after correct lever is pressed and reward is delivered. *Right*, No increase in firing rate after incorrect lever is pressed. Delivery of reward and sound of solenoid are absent during error trials (dopamine neuron from A10). Vertical scale 10 impulses/bin. Reproduced from Schultz et al. (1993) with permission from *The Journal of Neuroscience*.

from both sensory and reward representations to dopamine neurons is modifiable. (4) Some of these neurons have access to a representation of the expected time of reward delivery.

These data also show that simply being a predictor of reward is not sufficient for dopamine neuron responses to transfer. After training, as shown in Figure 2, the dopamine neuron response does not occur to the trigger light in the instructed spatial task, whereas it does occur to the trigger light in the spatial delayed response task. One difference between these tasks is that the trigger occurs at a consistent fixed delay in the instructed spatial task and at a randomly variable delay in the delayed response task (Fig. 2*C*).

Taken together, these data appear to present a number of complicated possibilities for what the output of these neurons represents and the dependence of such a representation on behavioral context. Below we present a framework for understanding these results in which sensory–sensory and sensory–reward prediction is subject to the same general learning principle.

THEORY

Prediction

One way for an animal to learn to make predictions is for it to have a system that reports on its current best guess, and to have learning be contingent on *errors* in this prediction. This is the underlying mechanism behind essentially all adaptation rules in engineering (Kalman, 1960; Widrow and Stearns, 1985) and some learning rules in psychology (Rescorla and Wagner, 1972; Dickinson, 1980).

Informational and structural requirements of a "prediction error" signal in the brain

The construction, delivery, and use of an error signal related to predictions about future stimuli would require the following: (1) access to a representation of the phenomenon to be predicted such as the amount of reward or food; (2) access to the current predictions so that they can be compared with the phenomenon to be predicted; (3) capacity to influence plasticity (directly or indirectly) in structures responsible for constructing the predictions; and (4) sufficiently wide broadcast of the error signal so that stimuli in different modalities can be used to make and respond to the predictions. These general requirements are met by a number of diffusely projecting systems, and we now consider how these systems could be involved in the construction and use of signals carrying information about predictions.

Predictive Hebbian learning: making and adapting predictions using diffuse projections

The proposed model for making, storing, and using predictions through diffuse ascending systems is summarized in Figure 4 (Quartz et al., 1992; Montague et al., 1993, 1995; Montague and Sejnowski, 1994; Montague, 1996).

Neuron P is a placeholder representing a small number of dopamine neurons that receive highly convergent input from both cortical representations $x(i, t)$ and inputs carrying information about rewarding and/or salient events in the world and within the organism $r(t)$, where i indexes cortical domains and t indexes time. Each cortical domain i is associated with weights $w(i, t)$ that characterize the strength of its influence on P at time t after a its onset. The output of P is widely divergent. The input from the cortex is shown as indirect—first synapsing in an intermediate layer. This is to emphasize that weight changes could take place anywhere along within the cortex or along the multiple pathways from the cortex to P, possibly including the amygdala.

The connections onto P are highly convergent. Neuron P collects this highly convergent input from cortical representations in the form:

$$\sum_i \dot{V}(i, t), \tag{1}$$

where $\dot{V}(i, t)$ is some representation of a temporal derivative of the net excitatory input to cortical domain i at time t and $V(i, t) = x(i, t)w(i, t)$. We use $\dot{V}(t) \equiv \Sigma_i V(i, t) - V(i, t - 1)$. P also receives input from representations of salient events in the world and within the organism through a signal labeled $r(t)$. The output of P $[\delta(t)]$ is taken as a sum of its net input and some basal activity $b(t)$:

$$\delta(t) = r(t) + \dot{V}(t) + b(t). \tag{2}$$

For simplicity, we let $b(t) = 0$, keeping in mind that the sign carried by $\delta(t)$ represents increases $[\delta(t) > 0]$ and decreases $[\delta(t) < 0]$ in the net excitatory drive to P about $b(t)$. If we let $V(t) = \Sigma_i V(i, t)$, the ongoing output of P $[\delta(t)]$ can be expressed as:

$$\delta(t) = r(t) + V(t) - V(t - 1). \tag{3}$$

Weight changes are specified according to the Hebbian correlation of the prediction error $\delta(t)$ (broadcast output of P) and the previous presynaptic activity (Rescorla and Wagner, 1972; Sutton and Barto, 1981, 1987, 1990; Klopf, 1982; Widrow and Stearns, 1985):

$$w(i, t - 1)_{\text{new}} = w(i, t - 1)_{\text{prev}} + \eta x(i, t - 1)\delta(t), \tag{4}$$

where $x(i, t - 1)$ represents presynaptic activity at connection i and time $t - 1$, η is a learning rate, and $w(i, t - 1)_{\text{prev}}$ is the previous value of the weight representing timestep $t - 1$. As shown in Figure 4, this model has a direct biological interpretation in terms of diffuse dopaminergic systems; however, this formula-

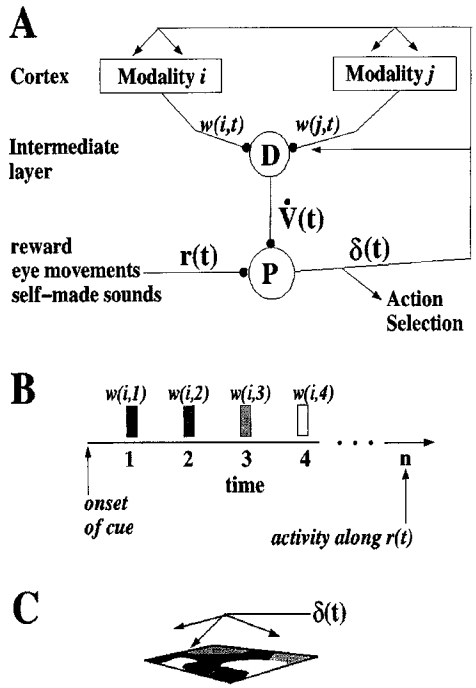

Figure 4. Making and using scalar predictions through convergence and divergence. *A, Modality i* and *Modality j* represent cortical regions. Neuron P collects highly convergent input from these cortical representations in the form $\Sigma_i \dot{V}(i, t)$, where $\dot{V}(i, t)$ is some representation of a temporal derivative of the net excitatory input to region i in the cortex. As indicated by the convergence through an intermediate region (neuron D), such temporal derivatives (transient responses) could be constructed at any point on the path from the cortex to the subcortical nucleus. We use $V(t) - V(t - 1)$ for $\dot{V}(i, t)$; however, other representations of a temporal derivative would suffice. The high degree of afferent convergence and efferent divergence permits P to output only a scalar value. P also receives input from representations of salient events in the world and within the organism through a signal labeled $r(t)$. This arrangement permits the linear output of P, $\delta(t) = r(t) + V(t) - V(t - 1)$, to act as a *prediction error* of future reward and expectations of reward (see text). Note that $\delta(t)$ is a signed quantity. We view this feature simply as increases and decreases of the output of P activity about some basal rate of firing that results in attendant increases and decreases in neuromodulator delivery about some ambient level. *B*, Representation of sensory stimuli through time. Illustration of serial compound stimulus described in the text. The onset of a sensory cue, say a green light, elicits multiple representations of the green light for a number of succeeding timesteps. Each timestep (delay after cue onset) is associated with an adaptable weight. At trial n, $r(t)$ becomes active because of the delivery of reward (juice). *C*, A simple interpretation of the temporal representation shown in *B*, the onset of the sensory cue activates distinct sets of neurons at timestep *1*, which results in a second group being activated at timestep *2*, and so on. In this manner, different synapses and cells are devoted to different timesteps; however, at any given timestep, the active cells/synapses represent green light from the point of view of the rest of the brain.

tion of the model also makes a direct connection with established computational theory. In particular, our formulation of the learning rule comes from the method of temporal differences (Sutton and Barto, 1987, 1990; Sutton, 1988). In temporal difference methods, the goal of learning is to make $V(t)$ anticipate the *sum* of future rewards $r(u)$, $u \geq t$ by encouraging predictions at successive time steps to be consistent. $\delta(t)$ in Equation 3 is a measure of the inconsistency, and the weight changes specified by Equation 4 encourage it to decrease. Further details of the rule are discussed in the Appendix. Predictions made by temporal difference methods are known to converge correctly under various conditions (Sutton, 1988), and they also lie at the heart of a method of learned optimizing control (Barto et al., 1989). The learning tasks addressed in this paper involve both classical and instrumental contingencies.

Representing a sensory stimulus through time

The occurrence of a sensory cue does not just predict that *that* reward will be delivered at some time in the future, it is known to specify *when* the reward is expected as well (Gallistel, 1990). This means that animals must have a representation of how long it has been since a sensory cue (like the light) was observed, and this information must be available at the level of the P.

We assume that the presentation of a sensory cue, say a light, initiates an exuberance of temporal representations and that the learning rule in Equation 4 selects the ones that are appropriate (Fig. 4*B*). We use the simplest form of such a representation: dividing the time interval after the stimulus into time steps and having a different component of x dedicated to each time step. This form of temporal representation is what Sutton and Barto (1990) call a complete serial-compound stimulus and is related to spectral timing model of Grossberg and Schmajuk (1989) in which a learning rule selects from a spectrum of timed processes. We do not propose a biological model of such a stimulus representation; however, in Figure 4*C* we illustrate one possible substrate.

Changing signal-to-noise ratios: translating prediction errors into decisions

In the absence of rewarding or reinforcing input, i.e., $r(t) = 0$, the fluctuating output of P reflects an ongoing comparison of $V(t - 1)$ and $V(t)$. Because these two quantities are predictions of *summed future rewards*, the difference between them indicates whether the future is expected to be more or less rewarding. Hence, through the weights that define $V(t)$, the output of P [$\delta(t)$] ranks transitions between patterns of activity in the cortex. In this manner, the weights $w(i, t)$ associated with the active cortical domains $x(i, t)$ act through the output of P to tag these *transitions* as "better than expected" [$\delta(t) > 0$] or "worse than expected" [$\delta(t) < 0$]. In our model of bumble-bee foraging based on the same theoretical framework, $\delta(t)$ was used in a similar manner to determine whether a bee should randomly reorient (Montague et al., 1995). Below, we use the same prediction error signal $\delta(t)$ to control behavior in a simple decision-making task. The same signal can be used to teach a system to take actions that are followed by rewards. This direct use of reinforcement signals in action choice may be a general phenomenon in a number of biological systems (Doya and Sejnowski, 1995).

RESULTS

Comparison of theory to physiological data

Training the model: learning with mistakes

Figure 5*A* shows the results of applying the model to the task given in Ljungberg et al. (1992), which is also similar to the spatial choice task in Figure 2. We address just the activity of the dopaminergic neurons that they recorded and do not address the process by which the monkey learns which *actions* to take to get reward. A light is presented at time $t = 41$, and a reward $r(t) = 1$ at timestep $t = 54$. As described above, the light is represented by a 20 component vector $x(i, t)$, where the activity of $x(i, t)$ for timestep k is 1 if $t = k$ and 0 otherwise. Figure 5*A* shows $\delta(t - 1)$ (output of neuron P) for three trials: before training, during training, and after significant training. Figure 5*B* shows $\delta(t - 1)$ for each timestep across the entire course of the experiment. In early trials (toward the *left* of Fig. 5*B*), the *prediction error* $\delta(t)$ is

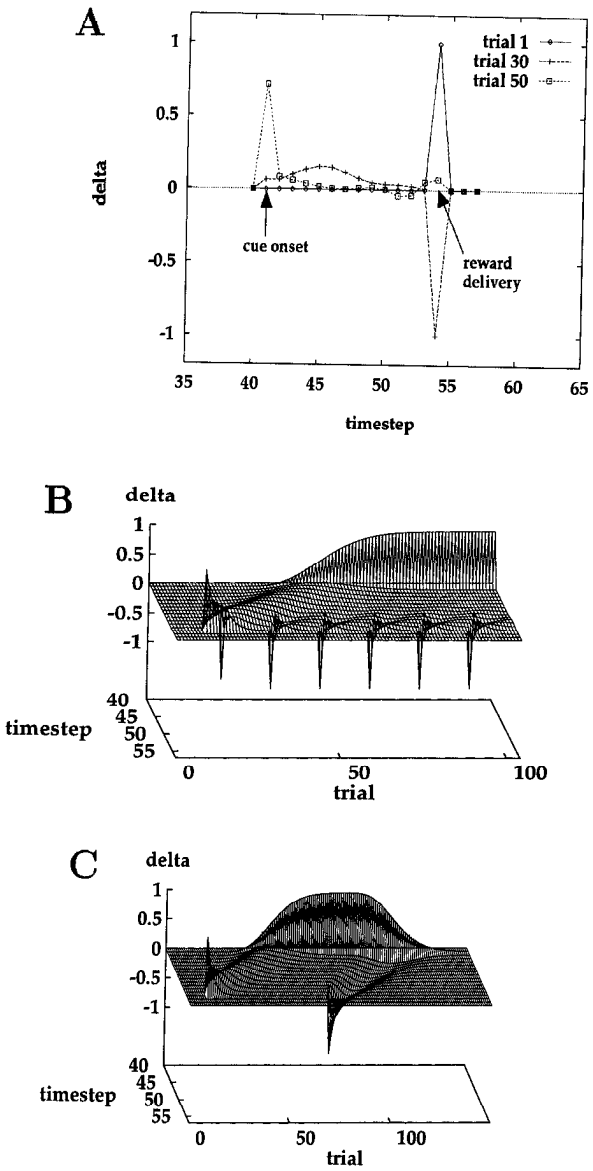

Figure 5. Model for mesolimbic dopamine cell activity during monkey conditioning. *A*, Plot of $\delta(t)$ (output of neuron P) over time for three trials during training. Each learning trial consisted of 120 timesteps. The model is presented with a single sensory cue at timestep 41 and reward $[r(t) = 1]$ at timestep 54. Initially (trial 1), the output of P $[\delta(t)]$ is large at the time that the reward is delivered ($t = 54$). During an intermediate trial (trial 30), the sensory cue is presented as before but the reward is withheld. At later trials (trial 50), the output of P $[\delta(t)]$ is large after the onset of the sensory cue and is near 0 at the delivery of reward. *B*, Entire time course of model responses. Sensory cue and reward delivery occur as in *A*, but training begins at trial 10. Over the course of ~60 trials, the largest change in the output of P shifts from timestep 54 to timestep 41. During intermediate trials, the prediction error is spread out through time so that, in the presence of some threshold for changes in firing rate, one would not necessarily expect to see an increase in firing rate moving back through time. Rather, in the case that we explore, the most noticeable firing rate changes in the mesolimbic dopamine neurons would appear *initially* at the presentation of reward and, after a number of trials, would appear locked to the presentation of the sensory stimulus. To simulate mistakes (wrong lever pressed), reward was withheld every 15 trials, resulting in a negative fluctuation in $\delta(t)$ (e.g., trial 30 illustrated in *A*). This negative fluctuation would be seen as a sharp decrease in the firing rate of the dopamine neuron. This means that *both during and after* learning, mistakes would be attended by decreases in firing at the time that the reward would have been delivered. This effect has been observed during learning (see Fig. 3). *C*, Extinction of response to the sensory cue. Model is trained as in *A* except that reward is always delivered until trial 70, after which reward is no longer delivered. As before, there is a negative fluctuation in $\delta(t)$ at the time that the reward would have been delivered, and by about trial 120, the response to the onset of the sensory cue has disappeared (learning rate in all panels: $\eta = 0.3$).

concentrated at the time steps when the reward is present. By the final trial, the *prediction error* is concentrated at the step when the light first comes on. Every 15 timesteps, reward was withheld, resulting in a large negative deflection in $\delta(t)$ that would be seen in the real experiment as a cessation in spike production.

Extinction of response to the sensory cue

Model is trained as in Figure 5*A* except that reward is always delivered until trial 70, after which reward is no longer delivered. As before, there is a negative fluctuation in $\delta(t)$ at the time that the reward would have been delivered and by about trial 120, the response to the onset of the sensory cue has disappeared (learning rate in all panels of Fig. 5: $\eta = 0.3$).

Instructed spatial task and delayed spatial task: the influence of temporal consistency

Figure 6 shows the training of the model in the presence of two predictive sensory cues. The initial sensory cue is presented at timestep 60 followed by presentation of the second sensory cue. In Figure 6*A*, this second cue occurs at a delay randomly varying from 9 to 11 timesteps. In Figure 6*B*, the second cue occurs *exactly* 10 timesteps after the initial cue. In both cases, the reward is presented at timestep 80 and lasts for 1 timestep. In Figure 6*A*, the model learns the magnitude and time of onset of the reward; however, it only partially discounts the onset of the second sensory cue. Note that throughout training, there remain fluctuations in $\delta(t)$ near the time of the onset of the second cue. This result captures the data in Figure 2*C*. This result changes if the second cue occurs at a consistent time after the initial cue. In this latter case, the model learns to discount the future onset of the second sensory cue and the reward consistent with the data in Figure 2*B*. This example suggests that the difference in dopamine neuron response in Figure 2*B* and *C*, depends on the consistency of the time of onset of the trigger light relative to the instruction light.

Influence of the temporal representation: response of the model in a noisy environment

One critical issue is the influence of other sensory cues that do not consistently predict the delivery of reward along $r(t)$. We show in Figure 7 the influence of the chosen temporal representation for sensory cues. In this figure, a sensory cue comes on and off (timestep 60), and its representation through time persists for 60 timesteps. The weights for each time step are initially randomized and are updated as before on each presentation of the cue (trial). Figure 7*A* shows $\delta(t)$, and Figure 7*B* shows the weights. The fluctuations in $\delta(t)$ rapidly decay to 0 except on the initial timestep where the fluctuation in $\delta(t)$ persists (>100 trials). Such persistence is attributable to the influence of the boundary conditions of the sensory representation as illustrated in Figure 7*B*.

The point of this example is to show how stimuli unrelated to reward delivery could influence the training of the model. If the temporal representation of a sensory stimulus does not overlap the period between the actual sensory predictor and the reward, then the learning rule averages away any initial weights because of the influence of the boundary of the representation. When these other stimuli overlap the period between the actual predictor and

←

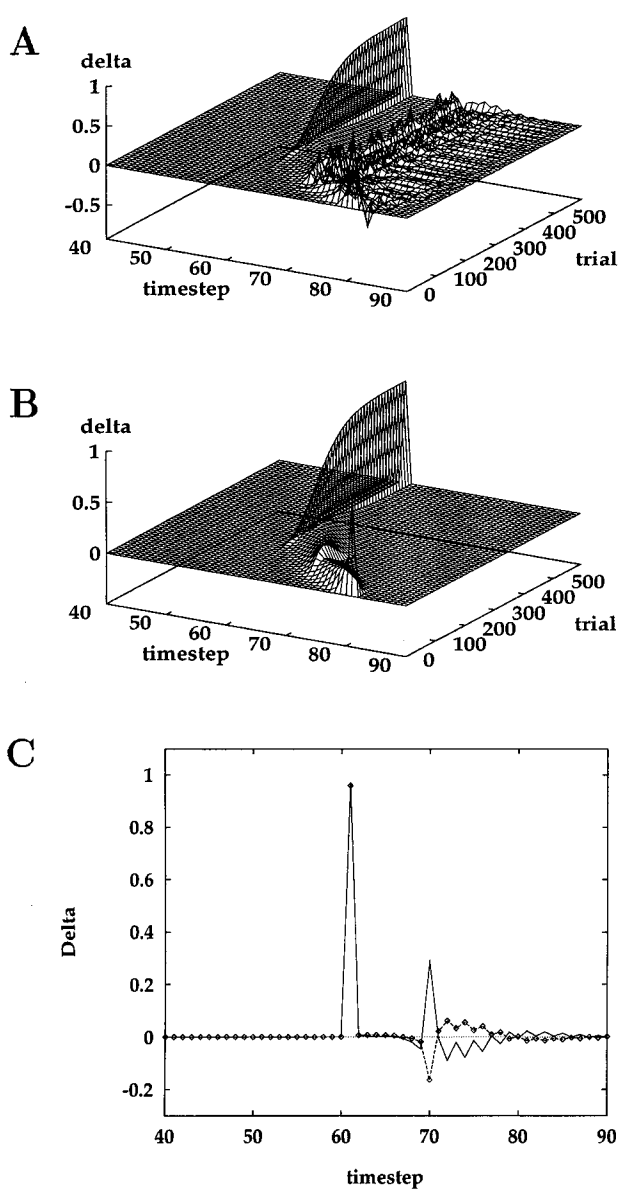

Figure 6. Response of model to two consistent predictors: instructed spatial task and delayed spatial task. *A* and *B* show plot of $\delta(t)$ (output of *P*) versus timestep and trial. *A*, Two sensory cues are presented followed by reward [$r(t) = 1$] at timestep 80. The first sensory cue is presented at timestep 60, and the second sensory cue is presented at a random delay varying from 9 to 11 timesteps later ($t = 69$–71). As before, $\delta(t)$ is large on the initial delivery of reward at $t = 80$. The model does not distinguish between fluctuations in $\delta(t)$ attributable to other sensory cues and rewarding input; hence, the weights develop so as to discount both. However, in this case, the second sensory cue does not occur at a predictable time so that $\delta(t)$ fluctuates near timestep 70 from trial to trial. After training, a histogram of the activity of *P* would show an increased activity near trial 70 that was more spread out through time than the response of *P* to the initial sensory cue. This example is analogous to the spatial delayed task. *B*, The model was trained as in *A* except that the second sensory cue was presented at exactly 10 timesteps after the first sensory cue. The weights develop so as to discount the occurrence of the second sensory cue and the reward delivery at $t = 80$. This example is analogous to the instructed spatial task. *C*, $\delta(t)$ (*Delta* in panel) from *A* shown for trials 500 (*line with diamonds*) and 501 (*solid line*). The positive fluctuations would tend to cause spikes and, therefore, contribute to a peristimulus histogram, whereas the negative fluctuations would be ignored if baseline firing rates were sufficiently low. The inconsistency of the relative time of presentation of the second sensory cue causes fluctuations $\delta(t)$ spread through time as shown (learning rate $\eta = 0.05$).

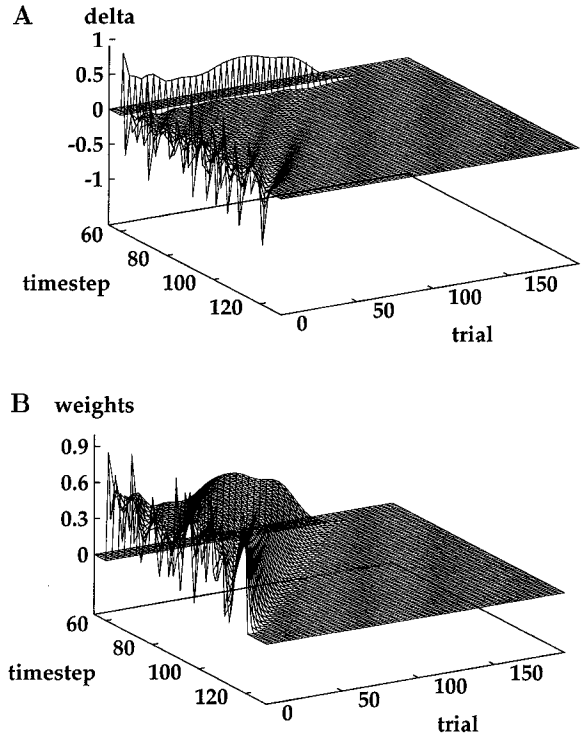

Figure 7. Response of model to unpredictive cues. Sensory cue comes on at timestep 60, goes off at timestep 61, and the representation of this event lasts for 59 succeeding timesteps. The initial weights for each time step are drawn from a uniform distribution on the interval (0,1) and are updated according to Equation 13. *A*, $\delta(t)$ as a function of time and presentation (trial). Except for the initial timestep when the sensory stimulus is presented, the fluctuations in $\delta(t)$ rapidly decay to 0. *B*, Weights associated with each timestep are shown as a function of the presentation (trial). The persistence of positive $\delta(t)$ at the onset of the stimulus is caused by the boundary conditions of the sensory representation that we use the learning rule in Equation 13. There are three effects illustrated in this example. (1) $\delta(t)$ is always negative just after the last timestep associated with a positive weight; therefore, the weights tend toward 0 beginning at the last timestep. This is seen as a ridge beginning at timestep 120 on trial 0 and progressing to timestep 61 around trial 120. (2) The learning rule implements a smoothness constraint and tends to drive the weights toward a stable point—an effect seen most clearly in the intermediate timesteps. (3) The weights for the initial timestep stay positive until the succeeding weight is driven to 0. This effect accounts for the persistent positive fluctuation in $\delta(t)$ shown in *A*.

the reward, they pick up weight changes and influence the value of $\delta(t)$, however, the strong averaging effect shown in the Figure quickly removes these changes during epochs when there is no overlap.

Physiological predictions

Figure 8 shows experimental consequences for the activity of the dopaminergic neurons in cases that have yet to be tested. In this example, a sensory cue consistently precedes the delivery of reward by 50 timesteps and the model trains as before. At trial 200, the time of reward delivery is reduced to 25 timesteps, i.e., reward delivery is twice as soon as would be expected. This change has three consequences: (1) in the first anomalous trial (200), $\delta(t)$ is *positive* at the new time of the reward; (2) $\delta(t)$ is *negative* at the old time; and (3) information that no reward is delivered at the original time takes longer than information about the new reward time to propagate back and affect the initial $\delta(t)$ in each trial. This latter consequence results in the transient elevation in the

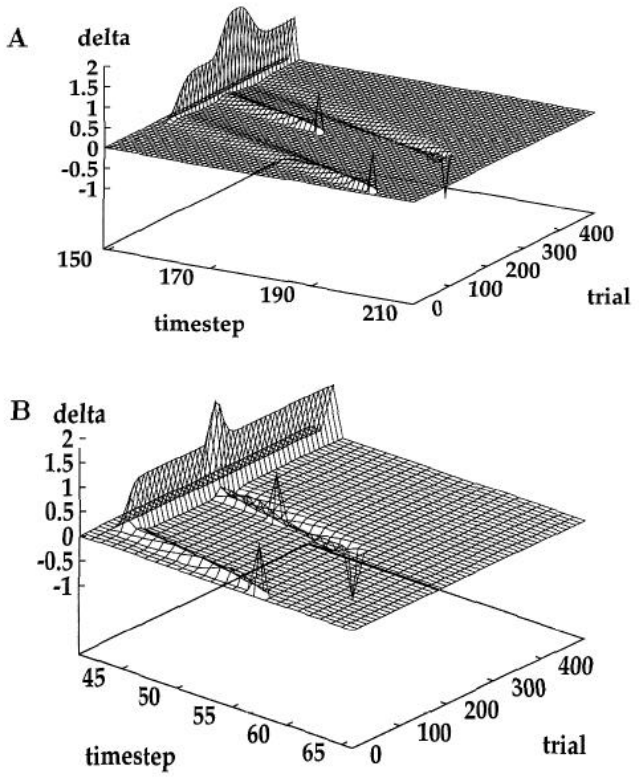

Figure 8. Physiological predictions. This figure shows some predictions of the model that depend on the complete serial-compound stimulus that we used for our temporal representation (see Fig. 4). The learning rate in both panels is $\eta = 0.3$. *A*, The model is trained as before with a sensory stimulus occurring at timestep 150 and the reward delivery 50 timesteps later at timestep 200. After trial 200, the reward is delivered 25 timesteps after the sensory stimulus instead of 50. At this switch of reward delivery, the model predicts that the dopamine cell would fire at the new delivery time of reward and would cease firing at the time that the reward had been delivered previously. The response to the sensory cue grows to twice its initial value and then decays back to its initial value with repeated presentations. This effect might not be noticed in practice because, for a fixed learning rate, it is proportional to the time difference between the old and new reward delivery times. This is illustrated in *B*. *B*, Model is trained exactly as in *A*; however, there are only 7 timesteps between the old and new reward delivery rather than 25.

stimulus-locked activity of the neurons seen between trials 250 and 350 in Figure 8*A*. Predictions 2 and 3 depend crucially on precise details of the representation of the stimulus over time (Fig. 8*B*). Prediction 1 should not depend on precise details of the stimulus representation.

Decision-making: predictions for human choice behavior

The preceding examples show how the model propagates information through time and how highly convergent descending connections can generate expectations about future rewards and predictions. In the above examples, the predictors of reward were extremely consistent and exhibited no variability in their temporal relation to the delivery of information about reward. We now ask how the model could use information about predictions of reward to bias actions. We also consider what happens if the delivery of reward is variable and depends on the history of the action choices. We show that under certain circumstances of reward delivery it would be difficult for an animal to maximize long-term reward delivery under our model of the influence of the diffuse dopamine systems on both synaptic plasticity and signal-to-noise ratios.

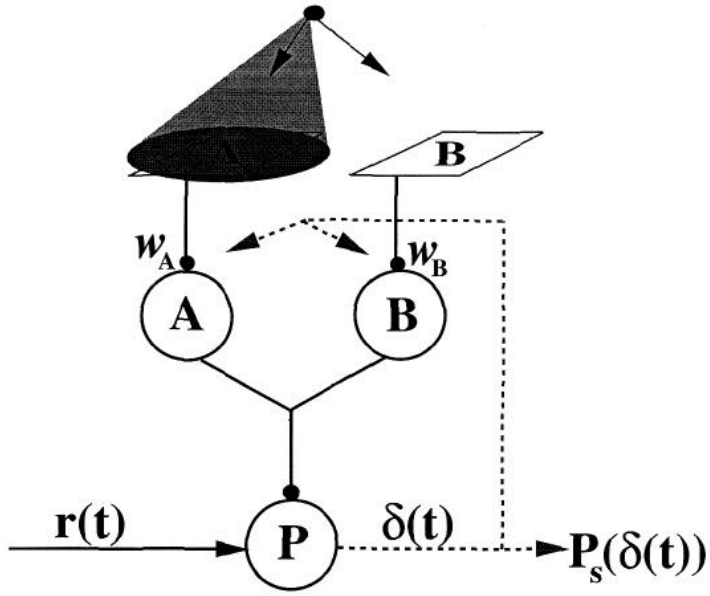

Figure 9. Card choice experiment. A card choice experiment involving two decks of cards was given to the network. The network made its card selections as follows: random transitions between decks (labeled *A* and *B*) were made to induce fluctuations in $\delta(t)$. Acting through P_s, the probability that the current deck is selected, the fluctuations in $\delta(t)$ acted to bias the choice of decks. After each selection of a card, a reward was delivered along $r(t)$ and weights (w_A, w_B) associated with each deck were updated according to Equation 8. The reward received at each card choice was a function of the fraction of the preceding 40 choices in which deck A was selected (reward functions shown in Fig. 8).

Card choice experiment

Figure 9 illustrates a card choice task given to the networks and humans (Egelman et al., 1995). The task is to select a card from one of the two decks of cards after which a reward is delivered along $r(t)$. As specified by the reward functions in Figure 10, the reward from both decks changes as a function of the percentage of choices from deck A. Notice that the reward functions cross at one point. To the right of this crossing, the reward function for deck B continues to grow and the reward function for deck A stays approximately the same. By design, it is suboptimal to choose cards in the ratio near the crossing point of the reward functions. The simple first-order nature of the model shows that it will pursue a hill-climbing strategy that will tend to get stuck at the crossing point.

The model and its behavior

The model is illustrated in Figure 9. The output of $P[\delta(t)]$ is again proportional to the sum of its inputs, so in this case we have:

$$\delta(t) = r(t) + V(t) - V(t-1), \quad (5)$$

$$V(t) = x_A(t)w_A(t) + x_B(t)w_B(t), \quad (6)$$

where $r(t)$ is 0 before a card is actually chosen. The selection of a deck was determined by Equation 7, where P_s is the probability of selecting the current deck. Fluctuations in δ were induced by allowing random transitions between the two alternatives. The model randomly chose one deck as a starting point and "looked back and forth" between decks, the fluctuations in $\delta(t)$ assigned a value to the transitions between choices, and P_s determined the probability that a given deck was selected after a transition. Weights (w_A and w_B) determined the sign and magnitude of

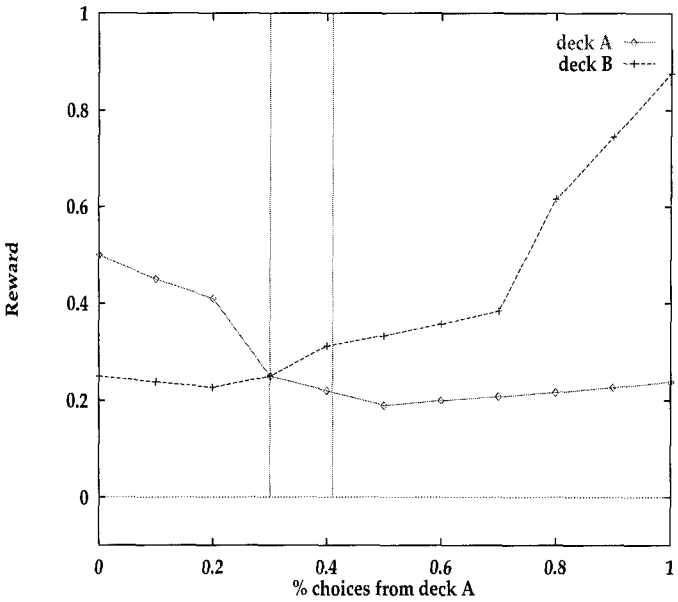

Figure 10. Reward functions and network performance. Reward functions for deck A and deck B as a function of the fraction of the last 40 cards chosen that were from deck A. For the network simulations, initial starting points were varied from 0.0 to 0.9 in steps of 0.1, and learning rates η were varied from 0.05 to 0.95 in steps of 0.05. m was varied over a range from 5 to 15. b was varied over a range from 0.0 to 1.0. For all of these conditions, the mean fraction of selections from deck A settled to values that fell between the vertical lines after ~200 iterations. Phase plots of the evolution of the weights w_A and w_B reveal that after 100 iterations, the weights settle into a stable basin of attraction.

fluctuations in $\delta(t)$, and thus influenced the choices between the decks:

$$P_s = \frac{1}{1 + \exp(m\delta(t) + b)}. \quad (7)$$

Permitting $\delta(t)$ to control noise levels at active target neurons allows the sign and magnitude of $\delta(t)$ to choose whether to "permit the action" or "wash it out" with increased noise levels. We do not attempt to specify the detailed dynamics of various elements of the model as a particular deck is chosen. Hence, some arbitrariness results in choosing the form of $\delta(t)$ during these events and we opt for a simplified version with $\gamma = 0$, i.e., $\delta(t) = r(t) - V(t-1)$, where $V(t-1) = x(i, t-1)w(i, t-1)$ with i indexing the selected deck. Justification for such simplification of $\delta(t)$ has been given in previous work (Montague et al., 1995). Using this form of $\delta(t)$, the weights associated with each deck were updated at each reward encounter (card choice) by Equation 8:

$$\Delta w(t) = \eta x(t)\delta(t). \quad (8)$$

The results for the network are shown as vertical lines showing the range of the mean fraction of deck A selections for all learning rates and initial starting positions. The parameters that determine the form of P_s (m and b) were varied over the ranges: m(0.1,5.0), b(0.0,15.0). The main influence of these parameters was to control the size of the basin of attraction for the sensory weights (w_A, w_B) and of course the dynamics of the approach to this basin. In these ranges, the network still converged on the range of 0.30 to 0.41 for the fraction of selections from deck A. In preliminary experiments, human subjects performed similar to the networks and tended to stick near the crossing points of the reward functions (Egelman et al., 1995).

DISCUSSION

We have proposed a particular relationship between the causes and effects of mesencephalic dopaminergic output on learning and behavioral control. The theory that we present accounts for a wide range of results in physiological recordings from dopamine neurons in behaving primates (Figs. 5, 6) and makes testable physiological predictions for future experiments (Fig. 8). The theory also makes strong predictions for a restricted class of decision-behaviors that is consistent with preliminary experiments in humans (Egelman et al., 1995).

Based on the success of these results, we postulate the following: *the fluctuating delivery of dopamine from the VTA to cortical and subcortical target structures in part delivers information about prediction errors between the expected amount of reward and the actual reward.* Under such a postulate, increases in dopamine release indicate that the current state is better than expected, decreases indicate that the current state is worse than expected, and the predictions (expectations) are represented in the pattern of weights that develop (see also Wickens and Kotter, 1995; Houk et al., 1995). In cases in which one sensory cue predicts another sensory cue as well as reward, the model develops weights that predict the time and magnitude of both future events, i.e., the model does not distinguish between stimulus–stimulus prediction and stimulus–reward prediction.

We have been very specific about how the dopamine neuron responses develop and the kind of information carried by fluctuations in their output (Fig. 4); however, we have not been specific about anatomical loci where weight changes may be stored. Ljungberg et al. (1992) and Schultz et al. (1993) report no difference between the dopaminergic cells in the VTA and those in the substantia nigra (although the frequencies are different). Given the involvement of the dorsal striatum in motor control, it is likely that there will be cells in the substantia nigra that are broadcasting $\delta(t)$ to influence the choice of actions. In addition to the nucleus accumbens (Koob and Bloom, 1988), the amygdala is a potential site for the weight changes that occur in the model (Gallagher and Holland, 1994).

Self-stimulation and the influence of agents affecting dopamine action

Artificial conditions induced by electrical stimulation or pharmacological agents that perturb the actions of dopamine offer insight into the impact of dopamine systems on behavior. Electrodes that artificially stimulate neuron P would generate and distribute a large positive prediction error to target structures innervated by P. If an animal controlled increases in the firing rate of P through a bar press, then the neural representation of the bar press would act as a predictor of future reward through dopamine release at targets. The learning rule would change the weights so that they predict the increase in the activity of P due to the initial rate of bar pressing. Through such learning, the initial change in the activity of P and attendant changes in dopamine release would then decrease for a given rate or pattern of bar pressing. Furthermore, if the output of P also influenced the learning of actions as we have suggested, then electrical stimulation would lead to the animal learning to press the bar more readily or more often.

Agents like cocaine that prolong the action of dopamine at target structures could have a number of effects on the model. One possible model of their action would be an increase in the

effective $\delta(t)$. A set of sensory cues associated with administration of such compounds would predict an effect of dopamine release attributable partially to the prolongation of the action of dopamine by cocaine, i.e., a fictitiously large prediction error. After training the model under these conditions, the sensory cues that predict the prolonged action of dopamine no longer predict the correct amount of dopamine in the absence of cocaine. In particular, presentation of these cues without cocaine administration cause a decrease in the firing of P and an attendant decrease in the current ongoing dopamine delivery, i.e., cues associated with drug administration that are not paired with the drug cause actual dopamine release to drop below the current baseline release. Other drugs of addiction that act partly through the dopaminergic system may also cause similar behavior on withdrawal.

Indeed, there is evidence that the dopamine concentration in the nucleus accumbens decreases after the cessation of chronic treatment with drugs of addiction such as morphine, ethanol, and cocaine (Acquas et al., 1991; Parsons et al., 1991; Rossetti et al., 1992; Diana et al., 1993). Unfortunately, there is contradictory evidence as to whether the dopamine concentration in the accumbens also decreases after withdrawal from amphetamines (Rossetti et al., 1992; Crippens and Robinson, 1994), although this would be expected from the model. The model suggests that, rather than reflecting direct pharmacological effects, physiological and behavioral effects that attend drug taking or drug removal may relate in a complicated manner to learning effects that are slow to reverse or to accrue.

Representing information without tonic firing

The data from monkey conditioning were one of the main constraints on our theory at the level of choosing a learning rule consistent with physiological findings. Ljungberg et al., (1992) recorded firing in dopaminergic areas during a reaction-time conditioning task and showed an apparent transfer in activity as a consequence of learning. In early trials, increased firing rates were locked to the delivery of a juice reward. Once the monkey had learned that a light stimulus reliably predicted the reward, the increased firing was locked to the presentation of the light. In a similar delayed response task, Schultz et al., (1993) noted that there was no sustained activity in the dopaminergic neurons during the time between the stimulus and the ultimate reward and concluded that this lack of firing "suggests that dopamine neurons do not encode representational processes, such as . . . expectation of external stimuli or reward." Under our theory, the lack of sustained firing is to be expected, and other temporal difference-based conditioning theories would make similar predictions about the transfer of firing (Moore et al., 1986). Theories that are not based on these principles, such as the attention-based account of Grossberg and Levine (1988), would have to explain these results in a different manner. The important point is that there may be many ways to represent information during delay periods that are not reflected simply as tonic firing.

Action choice in a simple decision-making task

The binary choice experiment has long been used to test how various aspects of reward schedules influence the choices made by animal or human subjects (see Bush and Mosteller, 1955; Gallistel, 1990). In experiments in which an animal is given multiple behavioral alternatives each of which yields rewards of various sizes or strengths, the animal tends to adjust its sampling of alternatives according to the relative rewards obtained from each. In contrast to these findings, it has been suggested that humans tend to maximize their returns in similar tasks and that matching may be restricted to less intelligent creatures. This latter view has been challenged by Herrnstein (1991) and others. The reward functions used in the card choice experiment are adaptations of similar reward functions used by Herrnstein (1991) in a task using human subjects (see also Herrnstein, 1961).

Using our simple "bottom-up" neural model of the potential influence of dopamine delivery on target neurons, we observed that the model behaved so as to match the relative rates of return from the two decks independent of starting position, learning rate, and noise level in the decision function P_s. The capacity of the model to demonstrate stable matching behavior depends on the fluctuations in $\delta(t)$ so that reductions in these fluctuations would influence expected behaviors. There are two ways to reduce the influence of $\delta(t)$ on the behavior of the model: (1) decrease the magnitude of the fluctuations in $\delta(t)$ that would slow down learning, and (2) decrease the effect of the transmitter at the target. Case 2 is equivalent to decreasing the magnitude of positive fluctuations in $\delta(t)$ and leaving the magnitude of negative fluctuations unchanged. In addition, a blunting of the influence of $\delta(t)$ would also be expected from a lesion of the VTA (see Wise, 1982); however, a complete lesion of the VTA appears to block reward-dependent learning completely, thus preventing solid conclusions about its significance.

To maximize long-term returns, a more rational agent than our network (say a human) should choose cards so that the percentage from deck A fluctuates around 0.8. In preliminary experiments, such a strategy is discovered only by a minority of the participants, and most tend to choose cards to match the relative rate of return from each deck (Egelman et al., 1995). Many explanations cast at a variety of levels have been offered to explain such matching behavior, and various strategies can be formulated to achieve optimal outcomes (von Neumann and Morgenstern, 1947; Bush and Mosteller, 1955; Luce and Raiffa, 1957). We of course do not attempt to improve or amend such efforts here. We note, instead, that our proposed model provides one possible bottom-up description for how diffuse systems could establish constraints that favored event matching while not excluding other more complicated reward-seeking strategies. This may explain why it is difficult for animals to maximize long-term rewards and why under appropriate circumstances they appear to be risk-averse (Luce and Raiffa, 1957; Harder and Real, 1987; Real, 1991).

Our theoretical framework shows how dopamine systems could respond to appropriate statistical structure in a task to influence behavior. Other diffuse systems also send projections down the spinal cord or to other systems that project down the spinal cord; therefore, the prediction error signals in our theory may be attended by more peripheral responses such as skin conductance changes. In this manner, the framework that we have presented may provide a starting point for explaining observations of skin conductance responses during various learning, recognition, and decision-making task (Tranel and Damasio, 1985; Bechara et al., 1994). It will be interesting to test human subjects on this decision-making task and compare their behavior to those of the model.

APPENDIX

At time t, an animal experiences stimuli that are represented by components of a vector $\vec{x}(t)$ with each component i dedicated to a separate stimulus or stimulus feature. The animal can also receive a scalar reward $r(t)$. Under temporal difference methods, the

computational goal of learning is to use the stimuli $\vec{x}(t)$ to predict a measure of the *discounted sum of future rewards* $V(t)$:

$$V(t) = \sum_{s>t} \gamma^{s-t-1} r(s). \quad (9)$$

$0 \leq \gamma \leq 1$ is called a discount factor that makes rewards that arrive sooner more important than ones that are delayed. This formulation, i.e., predicting the sum of future rewards, is an important advance over static conditioning models such as the Rescorla–Wagner rule (Rescorla and Wagner, 1972). An assumption of temporal difference methods is that the environment is Markovian: future rewards do not depend on past reward except through the current state $x(t)$. Hence, we denote $V(t)$ as $V(\vec{x})$. Given these assumptions, $V(\vec{x})$ satisfies the recursive relationship:

$$V(\vec{x}(t-1)) = r(t) + \gamma V(\vec{x}(t)) \quad (10)$$

or

$$\delta(t) = r(t) + \gamma V(\vec{x}(t)) - V(\vec{x}(t-1)), \quad (11)$$

where $\delta(t)$ is called the temporal difference error. In this paper, actual estimates of the predictions $\hat{V}(\vec{x}(t))$ are constructed as:

$$\hat{V}(\vec{x}(t)) = \vec{x}(t) \cdot \vec{w}(t). \quad (12)$$

Weight changes are specified as

$$w(i, t-1)_{\text{new}} = w(i, t-1)_{\text{prev}} + \eta x(i, t-1)\delta(t), \quad (13)$$

where $w(i, t)$ is the weight of timestep t for stimulus i in the estimate $\hat{V}(\vec{x})$. η is the learning rate. $\delta(t)$ takes on both positive and negative values corresponding to too small and too large predictions. The learning rule increases weights that produce positive fluctuations in $\delta(t)$ and decreases weights that produce negative fluctuations in $\delta(t)$, i.e., they change so as to make $\delta(t)$ small.

REFERENCES

Acquas E, Carboni E, Di Chiara G (1991) Profound depression of mesolimbic dopamine release after morphine withdrawal in dependent rats. Eur J Pharmacol 193:133–134.

Barto AG, Sutton RS, Watkins CJCH (1989) Technical Report 89–95. Amherst, MA: University of Massachusetts.

Bechara A, Damasio A, Damasio H, Anderson S (1994) Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 50:7–15.

Bernheimer H, Birkmayer W, Hornykiewicz O, Jellinger K, Seitelberger F (1973) Brain dopamine and the syndromes of Parkinson and Huntington: clinical, morphological and neurochemical correlations. J Neurol Sci 20:415–55.

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

Crippens D, Robinson TE (1994) Withdrawal from morphine or amphetamine: different effects on dopamine in the ventral-medial striatum studied with microdialysis. Brain Res 650:56–62.

Delong MR, Crutcher MD, Georgopoulos AP (1983) Relations between movement and single cell discharge in the substantia nigra of the behaving monkey. J Neurosci 3:1599–1606.

Diana M, Pistis M, Carboni S, Gessa GL, Rossetti ZL (1993) Profound decrement of mesolimbic dopaminergic neuronal activity during ethanol withdrawal syndrome in rats: electrophysiological and biochemical evidence. Proc Natl Acad Sci USA 90:7966–7969.

Dickinson A (1980) Contemporary animal learning theory. Cambridge: Cambridge UP.

Doya K, Sejnowski TJ (1995) A novel reinforcement model of birdsong vocalization learning. In: Advances in neural information processing systems, Vol 7. (Tesauro G, Touretzky D, Alspector J, eds). Cambridge: MIT, in press.

Egelman DM, Person C, Montague PR (1995) A predictive model for diffuse systems matches human choices in a simple decision-making task. Soc Neurosci Abstr 21:2087.

Fibiger HC, Phillips AG (1986) Reward, motivation, cognition: psychobiology of mesotelencephalic dopamine systems. In: Handbook of physiology. The nervous system. Intrinsic regulatory systems of the brain, Vol 4, pp 647–675. Bethesda: American Physiological Society.

Freeman AS, Bunney BS (1987) Activity of A9 and A10 dopaminergic neurons in unrestrained rats: further characterization and effects of cholecystokinin. Brain Res 405:46–55.

Gallagher M, Holland PC (1994) The amygdala complex: multiple roles in associative learning and attention. Proc Natl Acad Sci USA 91:11771–11776.

Gallistel CR (1990) The organization of learning. Cambridge: MIT.

Grossberg S, Levine DS (1987) Neural dynamics of attentionally modulated Pavlovian conditioning: blocking, interstimulus interval, and secondary reinforcement. Appl Optics 26:5015–5030.

Grossberg S, Schmajuk NA (1989) Neural dynamics of adaptive timing and temporal discrimination during associative learning. Neural Networks 2:79–102.

Harder LD, Real LA (1987) Why are bumble bees risk averse? Ecology 68:1104–1108.

Herrnstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav 4:267–272.

Herrnstein RJ (1991) Experiments on stable suboptimality in individual behavior. Am Econ Rev Papers Proc 83:360–364.

Houk JC, Adams JL, Barto AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds). Cambridge: MIT.

Kalman RE (1960) A new approach to linear filtering and prediction problems. J Basic Eng Trans ASME 82:35–45.

Klopf AH (1982) The hedonistic neuron. New York: Taylor and Francis.

Koob GF, Bloom FE (1988) Cellular and molecular mechanisms of drug dependence. Science 242:715–723.

Ljungberg T, Apicella P, Schultz W (1992) Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol 67:145–163.

Luce RD, Raiffa H (1957) Games and decisions: introduction and critical survey. New York: Wiley.

Mackintosh NJ (1983) Conditioning and associative learning. New York: Oxford UP.

Montague PR (1996) Biological substrates of predictive mechanisms in learning and action choice. In: Neural-network approaches to cognition: biobehavioral foundations (Donahoe PP, ed). New York: Elsevier Science, in press.

Montague PR, Sejnowksi TJ (1994) The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. Learn Memory 1:1–33.

Montague PR, Dayan P, Nowlan SJ, Sejnowski TJ (1993) Using aperiodic reinforcement for directed self-organization. In: Advances in neural information processing systems (Giles CL, Hanson SJ, Cowan JD eds). San Mateo, CA: Morgan Kaufmann.

Montague PR, Dayan P, Person C, Sejnowski TJ (1995) Bee foraging in uncertain environments using predictive Hebbian learning. Nature 376:725–728.

Moore JW, Desmond JE, Bethier NE, Blazis DE, Sutton RS, Barto AG (1986) Simulation of the classically-conditioned nictitating membrane response by a neuron-like adaptive element: response topography, neuronal firing, and interstimulus intervals. Behav Brain Res 12:143–154.

Oades RD, Halliday GM (1987) Ventral tegmental (A10) system: neurobiology. 1. Anatomy and connectivity. Brain Res 434:117–165.

Parsons LH, Smith AD, Justice Jr JB (1991) Basal extracellular dopamine is decreased in the rat nucleus accumbens during abstinence from chronic cocaine. Synapse 9:60–65.

Quartz SR, Dayan P, Montague PR, Sejnowski TJ (1992) Expectation learning in the brain using diffuse ascending projections. Soc Neurosci Abstr 18:1210.

Real LA (1991) Animal choice behavior and the evolution of cognitive architecture. Science 253:980–986.

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: the effectiveness of reinforcement and non-reinforcement. In: Classical conditioning. 2. Current research and theory (Black AH, Prokasy WF, eds), pp 64–69. New York: Appleton Century-Crofts.

Romo R, Schultz W (1990) Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. J Neurophysiol 63:592–606.

Rossetti ZL, Hmaidan Y, Gessa GL (1992) Marked inhibition of mesolimbic dopamine release: a common feature of ethanol, morphine, cocaine and amphetamine abstinence in rats. Eur J Pharmacol 221:227–234.

Sawaguchi T, Goldman-Rakic PS (1991) D1 dopamine receptors in prefrontal cortex: involvement in working memory. Science 251:947–950.

Schultz W (1992) Activity of dopamine neurons in the behaving primate. Semin Neurosci 4:129–138.

Schultz W, Romo R (1990) Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. J Neurophysiol 63:607–624.

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. J Neurosci 13:900–13.

Sutton RS (1988) Learning to predict by the methods of temporal difference. Machine Learning 3:9–44.

Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88:135–170.

Sutton RS, Barto AG (1987) A temporal-difference model of classical conditioning. Proceedings of the Ninth Annual Conference of the Cognitive Science Society. Seattle.

Sutton RS, Barto AG (1990) Time-derivative models of Pavlovian reinforcement. In: Learning and computational neuroscience (Gabriel M, Moore J, eds). Cambridge: MIT.

Tranel D, Damasio AR (1985) Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics. Science 228:1453–1454.

von Neumann J, Morgenstern O (1947) Theory of games and economic behavior. Princeton: Princeton UP.

Wickens J, Kotter R (1995) Cellular models of reinforcement. Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds). Cambridge: MIT.

Widrow B, Stearns SD (1985) Adaptive signal processing. Englewood Cliffs, NJ: Prentice-Hall.

Wise RA (1982) Neuroleptics and operant behavior: the anhedonia hypothesis. Behav Brain Sci 5:39.

Wise RA, Bozarth MA (1984) Brain reward circuitry: four circuit elements "wired" in apparent series. Brain Res Bull 12:203–208.