A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement

Article · July 1995

CITATIONS
587

READS
1,615

3 authors, including:

James C. Houk
Northwestern University
191 PUBLICATIONS **12,562** CITATIONS

SEE PROFILE

13 A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement

James C. Houk, James L. Adams, and Andrew G. Barto

INTRODUCTION

Reinforcement resulting from reward or punishment is important in shaping human and animal behavior, and there is considerable evidence that the dopamine (DA) system of the basal ganglia is a crucial brain mechanism for mediating learning by reinforcement (Beninger, 1983; Wise and Rompré, 1989; Wickens, 1990) Reinforcement is also emerging as a powerful strategy for confronting difficult problems in engineering control (Barto et al., 1990; Werbos, 1992). Although learning by reinforcement has many advantageous properties, it has an important limitation, called the credit assignment problem, that can be critical in learning applications (Minsky, 1963; Barto et al., 1983) In biological terms, this is the problem of getting reinforcement signals to the right synapses (spatial credit assignment) at the right time (temporal credit assignment) for them to be effective in guiding the learning process. Here we present a neural model that addresses the temporal credit assignment problem through the detection of events that predict subsequent reinforcement. We introduce this model based on the anatomy and physiology of the striosome compartments of the striatum (Gerfen et al 1987; Graybiel, 1991); and on the signaling properties of DA neurons (see chapter 12). We then give it a conceptual foundation that is based on the theory of adaptive critics described by Barto in chapter 11. Preliminary versions of the model have been published previously (Houk, 1992).

DOPAMINE NEURONS

There is now substantial evidence that DA neurons, located in the pars compacta of the substantia nigra and in the ventral tegmental area, play an essential role in both the primary reinforcement of behavior and in guiding preparatory behavior on the basis of the likelihood that the animal will subsequently receive reinforcement (Wickens, 1990; Apicella et al., 1991). In chapter 12, Schultz and colleagues review their findings

with microelectrode recordings from DA neurons in behaving monkeys showing that, at an early stage in learning a new behavioral task, these cells discharge in response to primary reinforcement. Later, as the animal learns the task, the cells begin to discharge in response to stimuli that regularly (or even probabilistically) precede the primary reinforcement and thus function as predictors of reinforcement. For example, Ljungberg et al. (1992) observed bursts of DA neuron discharge in response to any unexpected delivery of a drop of liquid reward in the initial phase of their experiment. Then the liquid reward was used as primary reinforcement in a task that required the monkey to reach and depress a lever when a small light was illuminated. As the monkey learned to respond to the light, bursts of DA discharge began to appear in response to the light, while responses to the liquid primary reinforcement progressively disappeared.

In another experiment probing the signaling properties of DA neurons, monkeys reached blindly into a box in search of a morsel of food, and DA bursts occurred whenever the fingers touched morsels of cookie or apple or a raisin (Romo and Schultz, 1990). In contrast, the neurons rarely responded to the touch of nonfood objects. Since this state of responsiveness served as the starting point for an experiment in which the animal was conditioned to an earlier event (the opening of the trap door to the food box) the authors treated contact with the food as primary reinforcement. However, these bursts are clearly in response to rather complex patterns of tactile input, as opposed to primary reinforcement produced by consuming the food. Here we postulate that the tactile responses to food objects are acquired, or secondary, reinforcers, because, during past experiences, they reliably predicted subsequent primary reinforcement obtained by food consumption and because they come to function as reinforcers themselves. In the next phase of the Romo and Schultz experiment, the door to the food box was left normally closed, and its abrupt opening signaled the availability of a morsel of food. During this phase the burst of DA discharge transferred from food contact to the opening of the food box that regularly preceded food contact. The noise and appearance of the trap door opening thus served as a predictor of food contact, which in turn predicted food consumption. These data suggest that DA discharge ratchets backward in time, in a sequence of familiar events, so as to respond to earlier and earlier predictors of reinforcement.

These intriguing results have both input and output implications that are each important to contemplate. In this chapter our main emphasis is on the input issues, i.e., what neural mechanisms enable DA neurons to fire in response to earlier and earlier predictors of reinforcement. We propose a model to explain how striosomal modules of the basal ganglia could predict future reinforcement in a recursive manner, and we show

how these signaling properties are analogous to those of an adaptive critic in the actor-critic architecture discussed by Barto in chapter 11. We also provide some discussion of the output issue, i.e., how might these signaling properties facilitate the control of the motor behaviors that ultimately secure the primary reinforcement. We will see how the system as a whole is potentially capable of addressing both the temporal and the spatial aspects of the credit assignment problem.

ORGANIZATION OF STRIOSOMAL MODULES

The input layer of the basal ganglia, the striatum, is divided into circumscribed regions called striosomes that are surrounded by matrix regions (Graybiel 1994; see also chapter 5). Both kinds of striatal region contain spiny neurons, so called because their processes are covered with dendritic spines that receive highly convergent input from the cerebral cortex and thalamus. However, the two regions differ in their chemical makeup and, most important, in the targets to which their neurons project (Gerfen et al., 1987; Graybiel, 1991). Spiny neurons in the striosomes project to DA neurons in the substantia nigra and ventral tegmental area, whereas spiny neurons in the matrix regions project to pallidal output neurons of the basal ganglia located in the internal division of the globus pallidus and in the pars reticulate of the substantia nigra. This chapter deals particularly with the striosomal spiny (SPs) neurons that project to DA neurons in the manner shown in figure 13.1.

The *solid black* arrow symbolizes the inhibitory GABAergic nature of the direct projection from SPs neurons to the DA neuron. The indirect projection involving a sideloop through the subthalamic (ST) nucleus instead has a net excitatory action. This ST sideloop is actually more complex than shown in figure 13.1, having a multisynaptic organization similar to the sideloops present in matrix circuits, as described by Graybiel and Kimura in chapter 5 and summarized by Houk in chapter 1 of this book. The simplified diagram will suffice for present purposes.

Spiny neurons in both striosome and matrix compartments have similar specializations in electrical properties (Kawaguchi et al., 1989) and receive organized, convergent input from widespread areas of the cerebral cortex (see chapter 5). The C's in figure 13.1 illustrate three of the thousands of cortical cells that send convergent input to the SPs neuron. As noted in chapter 1 and elsewhere, the neuronal architecture of the striatum is ideally suited for the recognition of complex patterns of cortical afference. We assume here that spiny neurons come to recognize complex contextual patterns through the reinforcing influence of the dopaminergic input to the striatum. In support of this hypothesis, there is growing evidence for dopamine-dependent plasticity in corti-

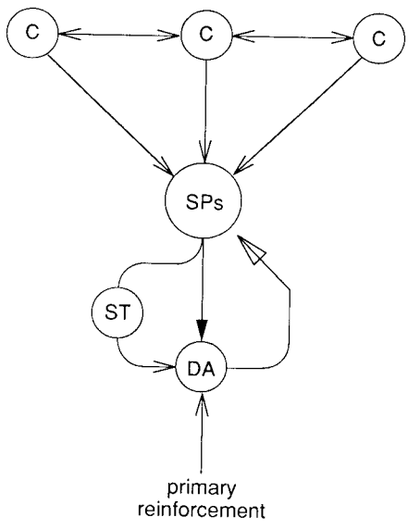

Figure 13.1 Striosomal module. See text for details. Open arrowheads signify net excitation, the black arrowhead in the direct SPs-to-DA projection signifies net inhibition, and the triangular arrowhead signifies neuromodulation. C, cerebral cortical columns; SPs, spiny neurons in striosomal compartments of the striatum; ST, subthalamic sideloop that includes neurons in the subthalamus; DA, dopamine neurons of the substantia nigra pars compacta.

costriatal synapses (see chapters 4 and 10 and below). The proposed specialization of SPs neurons for pattern recognition is assumed to allow them to detect particular contexts that are valid predictors of reinforcement. These responses would then be transmitted to DA neurons, which could explain how DA signaling predicts reinforcement. This hypothesis is elaborated below.

Another striking characteristic of the anatomy of striosomal modules is the projection that DA neurons make back to the same striatal zones that send them input (see figure 13.1). The reciprocal nature of projections between clusters of SPs and DA neurons has now been demonstrated in four laboratories (Gerfen et al., 1987; Jiménez-Castellanos and Graybiel, 1987; Selemon and Goldman-Rakic, 1990; Hedreen and DeLong, 1991); and can be considered a well-established aspect of striosomal modules. The model proposed here builds importantly on this characteristic, which we believe is responsible for the ability of DA neuron signaling to make progressively earlier predictions of reinforcement. As is explained more fully below, after an SPs neuron learns to fire in response to one context that predicts reinforcement, it can use this feedback pathway to reinforce itself for firing to an even earlier context that predicts reinforcement, which can then function as an antecedent secondary reinforcer. This anatomy gives rise to a recursive

feature, analogous to the adaptive critic's predictions of its own predictions, and we propose it as being essential for the resolution of the temporal credit assignment problem discussed in the opening paragraph.

A third input to DA neurons shown in figure 13.1 is labeled primary reinforcement. This connection is based on indirect argument, as opposed to direct anatomical demonstration. The technique of microdialysis has played an important role in demonstrating the likelihood that DA neurons receive signals from the lateral hypothalamus that are related to primary reinforcements of an appetitive nature (Wise and Bozarth, 1984; Hoebel et al., 1989).

MECHANISM OF RESPONSIVENESS TO PREDICTORS OF REINFORCEMENT

The organization of the striosomal module shown in figure 13.1 calls attention to three sources of input to DA neurons that might interact in the generation of DA firing patterns. Figure 13.2 shows hypothetical time courses of these signals in response to a "predictor of reinforcement" presented as a cortical input pattern that regularly precedes primary reinforcement, the latter signal being sent from the lateral hypothalamus.

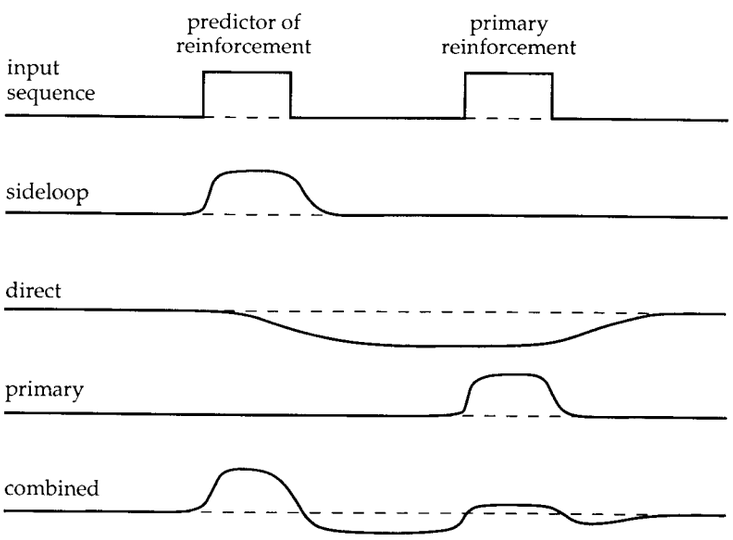

Figure 13.2 Hypothetical time course of signals generating dopamine neuron firing patterns. A pattern of excitatory signals which predicts reinforcement is delivered via projections from the cerebral cortical columns. The sideloop excitatory, direct inhibitory, and primary reinforcement signals converge on dopamine neurons, as shown in figure 13.1. The bottom trace shows the combination of those three signals.

Imagine, in analogy with the experiments of Schultz and colleagues, that the predictor of reinforcement is comprised of either the small light in the lever task, or the complex set of tactile stimuli that occur when the monkey contacts a piece of apple in the food box task. The primary reinforcement would be the liquid or food reward. We wish to explain why DA neurons respond to predictors of reinforcement and not to subsequent primary reinforcement, despite the latter being sent as excitatory input to DA neurons from the lateral hypothalamus.

In this section we assume that the SPs neuron has already learned, through a cellular mechanism that is described below, to fire a burst of discharge when the predicting context occurs. The traces in figure 13.2 labeled *sideloop* and *direct* respectively illustrate the postulated postsynaptic responses in the DA neuron produced by transmission of this burst through the excitatory ST sideloop and through the inhibitory direct projection. Our assumption that the excitation precedes the inhibition is in accord with electrophysiological observations showing that responses to electrical stimulation of the cortex or striatum evoke excitation followed by inhibition in substantia nigra (and globus pallidus) neurons (Kita and Kitai, 1991; Fujimoto and Kita 1992; Kita, 1992). The model, however, assumes a time course of inhibition slower than that observed in the electrophysiological experiments. These experiments thus far have demonstrated only a relatively rapid inhibition lasting about 25 ms and mediated by $GABA_A$ receptors. Although slow inhibition has not yet been demonstrated, its presence is anticipated based on the high density of $GABA_B$ receptors in the substantia nigra (Bowery et al., 1987; Martinelli et al., 1992). $GABA_A$ receptors act via G proteins to mediate slow inhibitory processes (lasting several hundred milliseconds) both in postsynaptic neurons and in presynaptic terminals (Isaacson et al., 1993).

The trace labeled *primary* in figure 13.2 shows an excitatory postsynaptic event postulated to occur in response to a primary reinforcement input from the lateral hypothalamus. If the latter were presented in isolation, the DA neuron would respond to it. However, if primary reinforcement is preceded by a predictive context that excites the SPs neuron, the complex sequence of postsynaptic events shown in the trace labeled *combined* would occur. The initial excitatory phase, owing to excitation transmitted through the sideloop, would fire the DA neuron. The subsequent inhibitory phase, transmitted through the direct pathway from the SPs neuron, might then largely cancel the excitatory potential produced by the primary reinforcement. The cancellation illustrated in figure 13.2 assumes that the inhibition is postsynaptic, as supported by monoclonal antibody staining (Martinelli et al., 1992). Cancellation might be even more complete if $GABA_B$ inhibition also occurred in the presynaptic terminals of the primary reinforcement input.

Thus far we have considered the relatively simple case of responding to a first-order predictor, i.e., a contextual event that is an immediate antecedent of primary reinforcement. In general, there may be a longer sequence of events and actions that ultimately leads to primary reinforcement. For example, the input sequence in figure 13.3 includes two contextual stimuli that function as successive predictors of reinforcement. Imagine that the later of the two predictors, C_a, represents the context that occurs when the monkey puts its hand into the food box and contacts a piece of apple. C_b represents an earlier, second-order, predictor, which might be the opening of the food box in the experiment of Schultz and colleagues. We assume the primary reinforcement, r, is the consumption of the apple.

The time plots below the input sequence illustrate the net postsynaptic activations in response to each event in the input sequence. (Note that this decomposition of input components is different from that used in figure 13.2.) The responses to C_a and to C_b each consist of initial excitatory phases followed by prolonged inhibitions, whereas the response to r is simply an excitation. The DA neuron would be expected to add together these components to generate its net output. It is apparent that the inhibitory phase of the response to C_b will cancel the excitatory phase of the response to C_a, and the inhibitory phase of the response to C_a will cancel the excitatory response to r. The bottom trace ignores the minor fluctuations of this summation process to illustrate that the net DA response, after all of these cancellations, would consist of a single excitatory phase produced by the earliest prediction of reinforcement, the context C_b. This is the pattern observed in recordings from DA neurons after DA signaling has transferred to the context associated with food box opening (see chapter 12).

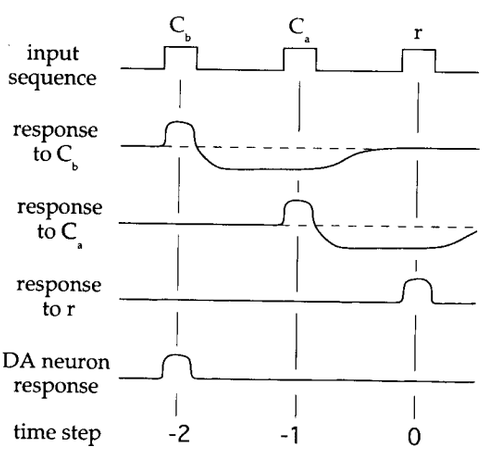

Figure 13.3 Dopamine (DA) neuron response to earlier predictors of reinforcement. See text for details. C_b and C_a, successive contextual stimuli; r, primary reinforcement.

CORRESPONDENCE WITH THE THEORY OF ADAPTIVE CRITICS

Barto in chapter 11 describes an adaptive critic as a device that learns to anticipate reinforcing events, pointing out that the effective reinforcement signals it generates can greatly enhance the performance of reinforcement learning systems. Influenced by the proposals of Klopf (1982), the adaptive critic was developed by Sutton (1988) and appeared as a neuronlike component of a control system (Barto et al., 1983) as well as a model of classical conditioning (Sutton and Barto 1981, 1990). Sutton (1988) used the term *temporal difference* (TD) algorithms for this general class of predictive mechanisms. There is a remarkable similarity between the discharge properties of DA neurons and the effective reinforcement signal generated by a TD algorithm under conditions representing the experimental situation of terminal primary reinforcement (see chapter 11.) The latter is the situation in which a sequence of contexts and associated actions must transpire before a primary reinforcing event ultimately occurs. Here we pursue this analogy by showing a particular correspondence between the signals that might by generated in striosomal modules and the terms in the TD equation for effective reinforcement.

Based on equation (10) from chapter (11), the effective reinforcement $\hat{r}_t$ at time step t is:

$$\hat{r}_{t-} = P_t - P_{t-1} + r_t \tag{1}$$

where P_t is the prediction at step t of future primary reinforcement, and P_{t-1} is the prediction at the previous time step. (We set the discount factor γ to 1 in order to simplify the discussion.) Equation (1) is a discrete-time version of the TD equation for effective reinforcement. To relate it to the activity of DA neurons, we suggest how these discrete time steps might relate to the continuous flow of real time by treating each discrete time step as the time of a salient event sensed by the animal. Various cellular events occur throughout the time intervals between these events.

Let us equate $\hat{r}_t$ with the discharge of the DA neuron in figure 13.1. Similarly, r can be equated with the primary reinforcement input to the DA neuron, which is zero except at the end of the requisite sequence of contexts and actions. At this point, it is natural to suggest that SPs neurons generate the predictions of reinforcement. P_t would then be transmitted by the excitatory sideloop and $-P_{t-1}$ by the direct, but slowly acting, inhibitory process. In essence, the slow inhibitory process functions as a kind of short-term memory of a negative image of the reinforcement that was predicted at the previous time step.

The time plots in figure 13.3 can be used to analyze $\hat{r}$ (DA neuron response) at each time step in the behavioral sequence from box opening (C_b), to food contact (C_a), to the food consumption that results in a

positive value for the primary reinforcement signal r_t at $t=0$. The earliest predictor C_b occurs at $t=-2$ and evokes a positive postsynaptic response that would represent the prediction P_{-2}. This is the only term that contributes to $\hat{r}_{-2}$, since there was no response of the SPs neurons at the previous time step $t=-3$. The response to C_b then goes through its negative phase, providing a negative trace of P_{-2} for the computation at $t=-1$. The predictor C_a then evokes a positive response for P_{-1}, but this is canceled by the negative trace of P_{-2}. Finally, at $t=0$, the primary reinforcement r_0 evokes a positive response that is canceled by the negative trace of P_{-1}.

The interpretation suggested by this comparison is that the burst discharges of SPs neurons might be thought of as predictions of subsequent reinforcement. In order for the model to conform with the specifics of the TD equation [equation (1)], these bursts would have to be generated at each stage of a behavioral sequence (though not necessarily by the same SPs neuron). This is one of the predictions of the present model that needs to be tested experimentally.

LEARNING TO PREDICT PRIMARY REINFORCEMENT

Earlier we described a mechanism capable of explaining the predictive responses of DA neurons that are observed after monkeys have learned a new behavioral task. This explanation was based on the assumption that, during the learning phase, SPs neurons projecting to the recorded DA neuron acquire an ability to respond to contextual inputs that are predictive of reinforcement. This is not a trivial learning task, since primary reinforcement is typically delayed, occurring a substantial time interval after the predictive contexts that need to be reinforced. Although synaptic plasticity has been demonstrated in striatal neurons (Calabresi et al., 1992), the issue of delayed reinforcement has not been addressed. In this section we present a cellular model of delayed DA-sensitive synaptic plasticity to explain how SPs neurons might learn to recognize contexts that are antecedent to DA reinforcement.

Lisman (1989) proposed a model of synaptic plasticity that is founded on the unique properties of a complex protein called calcium-calmodulin–dependent protein kinase II (CaM PK II). This molecule is a major component of the postsynaptic density (Kennedy et al., 1983), suggesting a general role in plasticity, and it is present in rather high concentrations in the striatum (Newman-Gage and Graybiel, 1988), suggesting a particular role in spiny neurons. Calmodulin (CaM) binding activates CaM PK II and subsequent autophosphorylation greatly prolongs the active state, presumably because it traps CaM in molecular pockets (Meyer et al., 1992; Schulman and Hanson, 1993). The prolonged active state then potentiates glutamate receptors so as to produce long-term potentiation (LTP) (McGlade-McCulloh et al., 1993). The model pro-

posed here incorporates this CaM trapping mechanism and also includes interactions between CaM PK II and the DA-stimulated intracellular signals studied extensively by Greengard and collaborators (Hemmings et al., 1987).

The flow diagram of figure 13.4 outlines the anticipated interactions between several intracellular signals in the spines of SPs neurons, and figure 13.5 illustrates how the dynamics of these interactions might increment the synaptic responsiveness to a context C_a that precedes reinforcement r. At the upper left in figure 13.4, glutamate (released at the terminals of cortical afferents in response to C_a) is shown to initiate a cascade of intracellular signaling. The three types of receptor for glutamate produce a mixture of membrane depolarization and an increase of intracellular Ca^{2+} (Bliss and Collingridge, 1993). The latter effect mediates plasticity, since it activates CaM, which activates CaM PK II, which then potentiates glutamate receptors. Bound CaM is nec-

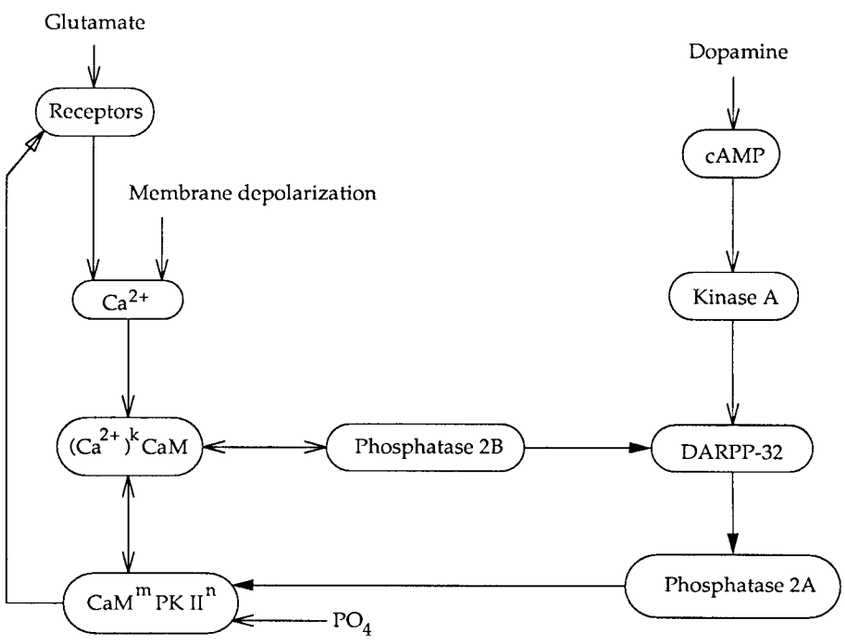

Figure 13.4 Cellular model of interactions which support learning earlier predictors of reinforcement. See description in text. Open arrowheads signify excitation or activation, black arrowheads signify inhibition or deactivation, and the triangular arrowhead signifies a possible potentiation or depression of receptors. Ca^{2+}, free calcium ions; $(Ca^{2+})^k$CaM, calcium-activated calmodulin with k (up to four) bound calcium ions; CaM^mPK II^n, activated Ca^{2+}-calmodulin–dependent protein kinase II with m pockets occupied by activated calmodulin molecules and n sites phosphorylated; PO_4, phosphate group; cAMP, the second messenger cylic adenosine monophosphate; DARPP-32, dopamine- and cAMP-regulated phosphoprotein.

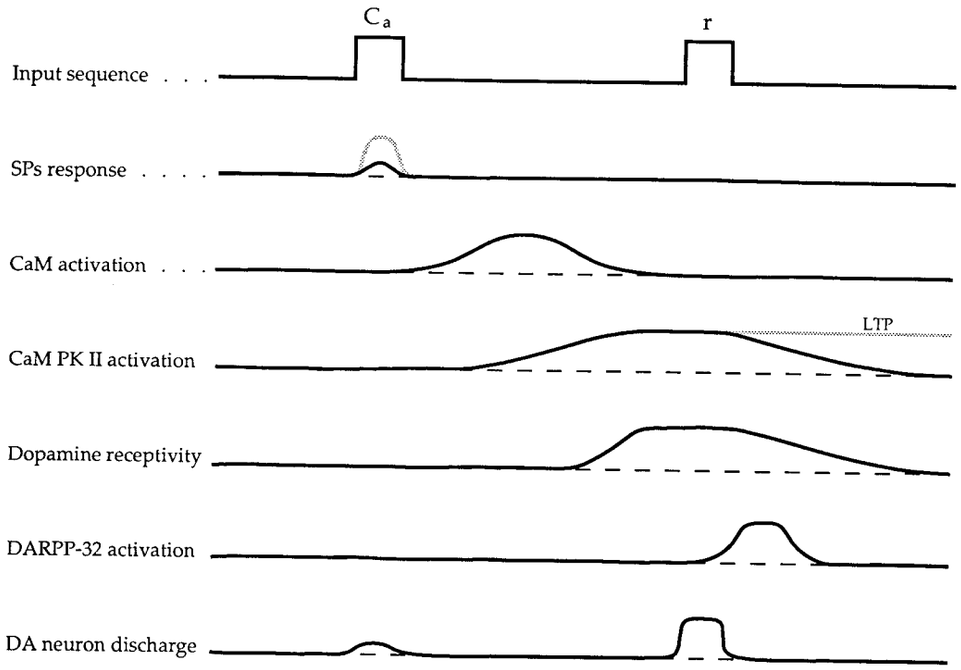

Figure 13.5 The interplay of intra- and extracellular events in creating long-term potentiation (LTP) of a predictor of reinforcement. See text and previous figures for details and abbreviations.

essary to activate its substrates and to initiate LTP (Mody et al., 1984). In figure 13.4, k in the expression $(Ca^{2+})^{k}CaM$ is the number of calcium ions bound to CaM, up to four; the delayed buildup of activated CaM in figure 13.5 is postulated to result from this multiple binding requirement. Although CaM has a high affinity for many of its substrates, including the phosphatase 2B shown in figure 13.4, it has a much lower affinity for binding to protein kinase II (PK II). This low affinity would be expected to raise the threshold and delay the time course for the activation of CaM PK II, as is illustrated by the solid trace of CaM PK II activation in figure 13.5.

The production of LTP requires an additional reaction, the autophosphorylation of CaM PK II. Autophosphorylation drastically alters the affinity of PK II for CaM, thus keeping this molecule activated for minutes as opposed to a fraction of a second in the dephosphorylated state (Schulman and Hanson, 1993). The dashed trace of CaM PK II activation labeled *LTP* in figure 13.5 illustrates the prolongation of activation that would be produced by autophosphorylation. We assume that this autophosphorylation step is normally counteracted by the presence of a phosphatase that drives the reaction toward dephosphor-

ylation. In figure 13.4 the enzyme that dephosphyhorylates PK II is assumed to be phosphatase 2A, because the usual enzyme, phosphatase 1, is apparently not present in striatal neurons (Nairn et al., 1988); furthermore, phosphatase 2A is more specific than, and has properties similar to, phosphatase 1 (Cohen, 1989). The several-minutes time course of CaM PK II activation (and LTP) produced by autophosphorylation could be extended to longer memories by protein synthesis mechanisms involving polyribosomes (Weiler and Greenough, 1993).

In our model, the conversion of CaM PK II into the autophosphorylated form, the critical step that initiates LTP, depends on the arrival of a properly timed DA reinforcement signal. DA works through DARPP-32, a *d*opamine- and-adenosine 3′:5′-cyclic monophosphate (*c*AMP)–*r*egulated *p*hospho*p*rotein extensively studied by Greengard's laboratory (Hemmings et al., 1987). This regulatory protein has its highest brain concentration in the striatum where it appears to be localized to the medium-sized spiny neurons. DA activates D1 receptors located on spines, which then activate cAMP, kinase A, and finally DARPP-32, as illustrated in figure 13.4. Activation of DARPP-32 inhibits phosphatase 2A, and thus can disinhibit the autophosphorylation of CaM PK II to initiate LTP, but this requires that the DA neuron fire during a brief period of DA receptivity.

The timing of DA neuron firing and DARPP-32 activation we assumed in the construction of figure 13.5 is appropriate to initiate LTP. Activated DARPP-32 inhibits phosphatase 2A, thus removing its inhibition of CaM PK II autophosphorylation. This, by itself, will not phosphorylate CaM PK II. Autophosphorylation occurs only under the condition that CaM PK II is already activated by CaM binding. Thus, the time course of CaM PK II activation shown in figure 13.5 is a permissive factor in defining the period of DA receptivity. Another condition for DA being effective relates to phosphatase 2B activity. The latter is rapidly activated and inactivated by transients in CaM activation, and, when activated, phosphatase 2B blocks the ability of DA to activate DARPP-32 (Halpain et al., 1990). Thus, the time course of CaM activity shown in figure 13.5 defines a prohibitive factor. The time course of DA receptivity shown in figure 13.5 reflects the combination of the permissive and prohibitive factors.

DA receptivity is analogous to the eligibility traces that have been invoked in certain computational models of reinforcement and classical conditioning (Klopt, 1982; Sutton and Barto, 1990; see also chapter 11). Like an eligibility trace, DA receptivity is a potential for reinforcement that becomes elevated after the occurrence of a synaptic input and decays slowly over time. However, unlike most eligibility traces used in network modeling, the elevation in DA receptivity does not begin immediately after a synaptic input; instead, it is postulated to be specifically delayed by the restrictive effects of phosphatase 2B activation.

As a consequence, spiny neuron synapses would tend to ignore contexts that precede reinforcement by very short time intervals. Instead, SPs neurons would preferentially learn antecedent contexts that precede reinforcement by a longer time interval, or, more precisely, by a range of longer time intervals as determined by the time course of the delayed DA receptivity.

While there are insufficient data on the time course of intracellular signaling to specify the preferential time interval discussed in the preceding paragraph, we postulate that it should be of the order of a few hundred milliseconds. This postulate is based on a requirement of the model, namely that the preferred interval for delayed reinforcement should not exceed the duration of the slow inhibitory event illustrated in figure 13.2. This is because an acquired response to a predictor needs to evoke an inhibition of sufficient duration to cancel a subsequent response to primary reinforcement, as was discussed earlier.

LEARNING EARLIER PREDICTORS OF REINFORCEMENT

The delayed receptivity of SPs neuron spines to DA reinforcement discussed in the previous section could provide an effective mechanism for learning a first-order predictor that occurs reliably at a fixed time interval prior to primary reinforcement. The predictive context would have to precede reinforcement by a few hundred milliseconds, which is a relatively short time interval on the scale of behavior. Learning to respond to higher-order predictors that may occur at longer and more variable delays is a more difficult problem. In this section, we explore the learning properties of an entire striosomal module when it contains an embedded SPs neuron with delayed DA receptivity. We show that these modules have an emergent property, a recursive capacity for learning to recognize a sequence of contextual events that are predictive of reinforcement. In the following section we discuss the similarity of this mechanism to the operations performed by an adaptive critic.

We have illustrated above how a context C_b, that reliably precedes a context C_a, that reliably precedes primary reinforcement r, could elicit discharge in the DA neurons of a striosomal module at the early time step when C_b occurs (see figure 13.3). There we assumed that the SPs neuron had already learned to fire in response to these contexts, whereas in this section we address the learning process itself. In particular, we describe how the DA released by a response to C_a could serve as a secondary reinforcer that trains the SPs neuron to recognize the context C_b that is an antecedent predictor of reinforcement. We then generalize this result to suggest how these modules could ultimately learn to recognize long chains of sequential events that lead to primary reinforcement.

The upper time plot in figure 13.6 shows the stimulus sequence, from C_b (the opening of the food box), which precedes the occurrence of C_a (the contact of the fingers with a piece of apple), which precedes the occurrence of the consumption of the piece of apple (the primary reinforcement r). We start with a state in which the SPs neuron has a strong response to C_a, owing to a previous history of primary reinforcement r, and an unreinforced, weak response to C_b, as shown by the second trace in figure 13.6. The weak response to C_b leaves in its wake a rising and falling wave of receptivity to DA in the spines that were excited by C_b. This is followed by a strong response to C_a, which gives rise to another wave of receptivity to DA, but this time in the spines that were excited by C_a. The DA response trace shows the net effect of these SPs neuron responses and the response to r on DA neuron discharge. Referring back to the examples of figures 13.2 and 13.3, one can mentally

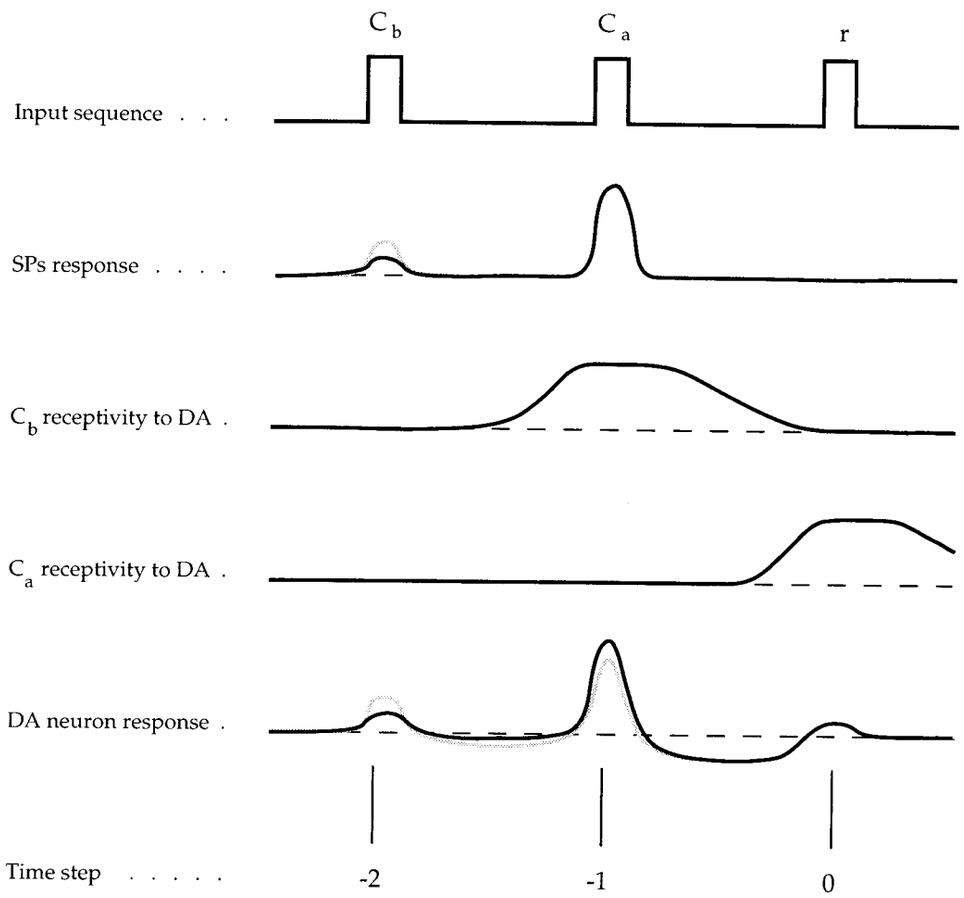

Figure 13.6 Learning a sequence of predictors of reinforcement. See text for discussion. Abbreviations as given for figures 13.1 and 13.3.

combine the different components of response to see how they would give rise to the DA response trace illustrated in figure 13.6. The salient feature of this trace, ignoring the smaller fluctuations, is the burst of discharge that occurs at the -1 time step.

Now we can contemplate how these several events that occur during a single trial would interact to produce learning. Based on the cellular model of plasticity discussed above, there would have to be a coincidence of DA release and of receptivity to DA in order to instantiate a change in synaptic strength. Figure 13.6 shows that the required coincidence occurs at $t=-1$ in those spines that responded to C_b at $t=-2$, and this instantiates an increase in synaptic strength in the C_b-responsive spines. The stippled traces in figure 13.6 show how this change in synaptic strength would affect responding in a subsequent trial with the same input sequence. The increased SPs neuron response to C_b is a direct result of enhanced transmission at the synapses responding to C_b. The response to C_a is unaltered in the SPs neuron, but is depressed in the DA neuron, as a consequence of the delayed inhibitory component of the enhanced response to C_b. The striosomal module as a whole thus acquires a stronger response to the earlier context and a weaker response to the later one. Over the course of many such trials, the context C_b would come to evoke an SPs neuron burst as strong as that shown in the original response to C_a. At this point, the burst of DA discharge, the earliest prediction of reinforcement, will have moved to the $t=-2$ time step, and the DA response to C_a will be completely inhibited. Note that responses to later contexts are not forgotten; instead, we predict that they are simply canceled by the delayed inhibitory components of responses to earlier contexts.

Having thus acquired this new antecedent of reinforcement, the striosomal module can then progress through yet another cycle of learning an even earlier predictor of reinforcement. Given the recursive nature of this process, striosomal modules might, through extended experiences, become capable of detecting contextual stimuli that occur at very long time periods in advance of a primary reinforcement. It is difficult to anticipate what the ultimate limit of this process might be. However, one might expect that the reliability of the prediction of reinforcement might progressively decrease at earlier times, thus diminishing the likelihood of finding a new context that predicts the DA response with sufficient regularity to become established as an antecedent predictor.

RELATION TO THE ACTOR-CRITIC ARCHITECTURE

The model of a striosomal module described in this chapter fulfills the main functions of the adaptive critic in the actor-critic architecture described by Barto in chapter 11. This architecture is an effective way of implementing a reinforcement learning system and is currently being studied by engineers and computer scientists as an approach to solving

difficult nonlinear control problems (Barto et al., 1983, 1990). The basic idea is to let predictions of reinforcement, which are generated by an adaptive critic, serve as surrogate (or secondary) reinforcers for controlling an actor, which is the system that generates the command signals that control actions. In the present section, we use this architecture as a framework for exploring how two kinds of information-processing module in the basal ganglia might function like interacting critics and actors in the control of motor behavior.

Earlier we reviewed the contrasting connectivity of spiny neurons in striosome and matrix compartments of the basal ganglia. As illustrated in figure 13.7, SPs neurons project to DA neurons and matrix spiny (SPm) neurons instead project to the pallidal (PD) output stage of the basal ganglia. Figure 13.7 indicates how this and other anatomical features serve to define matrix modules, which are partly analogous to the striosomal modules discussed above. Like the SPs neurons, SPm neurons have separate pathways for transmitting excitatory and inhibitory inputs to their target neurons, in this case the PD neurons. The descending connections of PD neurons are omitted in figure 13.7 to focus on the more prominent ascending pathways to columns of frontal cortical neurons (*F*), via specific divisions of the thalamus (*T*). We assume

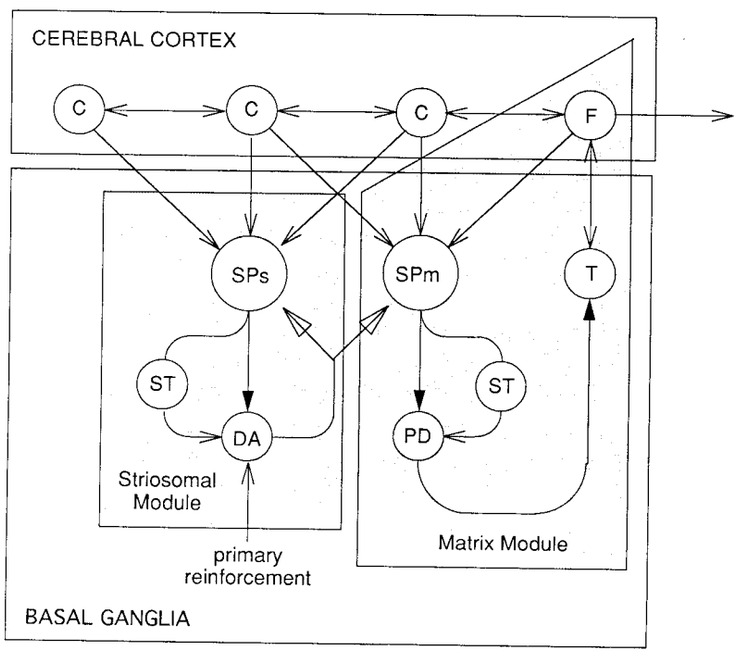

Figure 13.7 Modular organization of the basal ganglia including both striosomal and matrix modules. See figure 13.1 for the significance of the arrowheads and abbreviations for the striosomal module. Additional abbreviations: F, columns in frontal cortex; SPm, spiny neurons in the matrix compartment of the striatum; PD, pallidal neurons; T, thalamic neurons.

that the signals in F neurons function as the outputs of matrix modules. A specific hypothesis about the information-processing operations that go on in matrix modules is summarized in chapter 1 of this book.

Let us assume that striosomal molecules function as adaptive critics. In addition, we assume in figure 13.7 that the DA innervation of the striatum is sufficiently diffuse for the DA signal generated by a striosomal module to diverge to adjacent matrix modules. (This assumption is reconsidered in the next section.) Although only a single matrix module is shown, 95% of the spiny neurons are in matrix zones (Wilson, 1990), suggesting a ratio of about 20 matrix modules to each striosomal module. On this basis, one can justify divergence to the set of competitively interacting SPm units illustrated in figure 11.5 of chapter 11. Similarly, we postulate here that the groups of matrix modules innervated by a given striosomal module interact competitively with one another. It is then natural to assume that these groups of matrix modules function like the actor in the actor-critic architecture, thus generating signals that command various actions. (A more realistic assumption is discussed below.) Correspondingly, the branching of the DA fiber illustrated in figure 13.7 to both SPm and SPs neurons would provide the requisite coupling between a striosomal module functioning as an adaptive critic and surrounding matrix modules functioning as an actor. The predictions of reinforcement generated by the striosomal module would serve as secondary reinforcers for training the matrix modules. Under this training influence, the matrix modules would learn to use the convergent cortical input onto SPm neurons as a basis for generating action commands in sets of F neurons (see figure 13.7). The DA signals would tend to shape these commands into ones that regularly lead to the achievement of primary reinforcement.

Although the analogy proposed here between the basal ganglia and the actor-critic architecture is clearly a simplified abstraction of the actual anatomy and physiology of this part of the brain, it helps to translate into biological terms certain principles and concepts that may be important in basal ganglia function. As an example, we will use it to explain how striosomal modules might help to resolve the temporal credit assignment problem raised in the introductory paragraph. The temporal credit assignment problem arises because the process of adjusting a module's behavior on the basis of reinforcement feedback has to accommodate the fact that consequences are delayed and depend on the environmental conditions in which an action was made as well as on other actions executed before and after the action in question. As a result, both the nature of the consequences and their time courses are both delayed and highly variable. If an event occurs that has primary reinforcing significance, it is not easy to determine which elements of the preceding activity deserve the credit (or the blame in the case of an aversive event) for causing this event (the credit assignment problem alluded to above). The adaptive critic addresses this problem by learning

to predict, or anticipate, primary reinforcement over long time periods of variable and uncertain duration. The critic's predictions contribute to secondary, or acquired, reinforcement signals that provide *immediate* feedback about *anticipated* consequences of current activity (where "immediate" means after a relatively small and relatively fixed delay). Under the influence of these acquired reinforcement signals, which we postulate to be the signals produced by DA neurons, the matrix modules can learn to exert actions and thus influence future sensory input in desirable ways using synaptic modification rules that do not need to accommodate the highly variable temporal relationship between matrix activity and primary reinforcement.

MORE REALISTIC ASSUMPTIONS

While the model described in the previous section is attractive for its simplicity, some of its assumptions ignore the known anatomy and physiology of the system. Here we call attention to a few of these discrepancies, and discuss potential consequences of considering a more realistic model.

In the previous section we discussed the role of the basal ganglia as if there were only one striosomal module, computing a single prediction of reinforcement, and broadcasting that prediction broadly to all of the matrix modules. This most certainly is a gross oversimplification. Instead, we know that there are many, spatially segregated striosomes, each projecting to a somewhat separate cluster of DA neurons in a relatively topographic manner (Gerfen et al., 1987; Jiménez-Castellanos and Graybiel, 1987; Selemon and Goldman-Rakic, 1990; Hedreen and DeLong, 1991). Considering that each cluster of DA neurons is likely to receive somewhat different primary reinforcement inputs, the different modules associated with each striosome should become involved in predicting different qualities of primary reinforcement. Furthermore, since each striosome contains a large number of SPs neurons, each receiving a somewhat different constellation of afference, one would expect that different modules belonging to a given striosome would find different bases for predicting their special quality of primary reinforcement.

These anatomical considerations suggest considerable opportunity for generating a large diversity of secondary reinforcers that would then be available for training different groups of matrix modules. Why then have Schultz and colleagues (see chapter 12) observed such a high degree of homogeneity in the responses of the DA neurons that they have sampled? Perhaps this results, at least in part, from the fact that their animals are engaged in learning just a single behavioral task over the period during which the population is being sampled. It would be interesting to train the animals to perform concurrently several behavioral tasks, to see if different DA neurons specialize to predict in differ-

ent ways. Concurrent learning of several tasks, some with a common primary reinforcement and others with different ones, probably reflects more accurately the experiences of the animal in the wild. Certainly this situation presents a much more challenging, and more interesting, computational problem, appropriately suitable for the parallel architecture of the brain.

The discussion in the preceding section treated the outputs of matrix modules as if they were signals that command specific actions. In contrast, single-unit recordings from frontal cortical F neurons, the outputs of matrix modules, indicate that these signals are not immediate commands, but rather signify higher-order properties (Goldman-Rakic, 1987; Schultz and Romo, 1992). Looking at these properties from a motor perspective, they would appear to represent plans that might then organize other systems to generate the actual command signals for actions. If one instead looks at these signals from a sensory perspective, they appear to signal complex contexts that could indeed be useful in the formulation and implementation of plans and actions. Houk and Wise (1993, 1994) adopted the latter view of matrix module function and suggested mechanisms whereby salient contexts might be detected and registered into working memory for subsequent use by motor program generators in the cerebellum. Salient contexts were postulated to include the state of the organism, the desirability of the action, the actions planned in the near future, the location of targets of action, and sensory inputs that both select and trigger motor programs.

The linkages between F signals and actual motor commands is apparently quite flexible, which gives rise to an additional source of uncertainty with which the adaptive critics would have to contend. These concepts do not invalidate the actor-critic model of the basal ganglia suggested earlier. Instead, they cast this model within the perspective of more flexible, complex, and seemingly more powerful control options.

SUMMARY

The model presented here explains how dopamine (DA) neurons in the basal ganglia might acquire their ability to predict reinforcement and how outputs from these neurons might then be used to reinforce behaviors that lead to primary reinforcement. DA neurons are embedded in striosomal modules that include reciprocal connections with spiny neurons in the striatum. We propose a cellular learning rule whereby spiny neurons are trained by their DA input to detect contexts that precede reinforcement by a short time interval. Striosomal spiny neurons then use these acquired responses to control their own DA input. Through this recursive mechanism, DA neurons could learn to detect earlier and earlier predictors of reinforcement. DA signals also diverge

to reinforce spiny neurons in matrix modules, training the latter to detect and register contexts that are useful in planning and controlling motor behavior. The proposed scheme has interesting parallels with an actor-critic architecture that has been used to solve difficult engineering control problems.

ACKNOWLEDGMENTS

The authors are grateful to Richard S. Sutton for helping us relate the anatomy of striosomal modules to the mathematical operations performed by the temporal differences algorithm. This work was supported by a contract from the Office of Naval Research (N00014-88-K-0339).

REFERENCES

Apicella, P., Ljungberg, T., Scarnati, E., and Schultz, W. (1991) Responses to reward in monkey dorsal and ventral striatum. *Exp. Brain Res.* 85:491–500.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983) Neuronlike elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cyber.* 13:835–846.

Barto, A. G., Sutton, R. S., Watkins, and C. J. C. H. (1990) Learning and sequential decision making. In M. Gabriel and J. Moore (eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks.* Cambridge, Mass.: MIT Press, pp. 539–602.

Beninger, R. J. (1983) The role of dopamine in locomotor activity and learning. *Brain Res.* 287:173–196.

Bliss, T. V. P., and Collingridge, G. L. (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361:31–39.

Bowery, N. G., Hudson, A. L., and Price, G. W. (1987) GABA-A and GABA-B receptor site distribution in the rat central nervous system. *Neuroscience* 20:365–383.

Calabresi, P., Pisani, A., Mercuri, N. B., and Bernardi, G. (1992) Long-term potentiation in the striatum is unmasked by removing the voltage-dependent magnesium block of NMDA receptor channels. *Eur. J. Neurosci* 4:929–935.

Cohen, P. (1989) The structure and regulation of protein phosphatases. *Annu. Rev. Biochem* 58:453–508.

Fujimoto, K., and Kita, H. (1992) Responses of rat substantia nigra pars reticulata units to cortical stimulation. *Neurosci. Lett.* 142:105–109.

Gerfen, C. R., Herkenham, M., and Thibault, J. (1987) The neostriatal mosaic. II. Patch- and matrix-directed mesostriatal dopaminergic and non-dopaminergic systems. *J. Neurosci* 7:3935–3944.

Goldman-Rakic, P. S. (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum and V. B. Mountcastle (eds.), *Handbook of Physiology. The Nervous System V,* Part 1. Bethesda, Md., American Physiological Society, pp. 373–417.

Graybiel, A. M. (1991) Basal ganglia—input, neural activity, and relation to the cortex. *Curr. Opin. Neurobiol.* 1:644–651.

Halpain, S., Girault, J-A., and Greengard, P. (1990) Activation of NMDA receptors induces dephosphorylation of DARPP-32 in rat striatal slices. *Nature* 343:369–372.

Hedreen, J. C., and DeLong, M. R. (1991) Organization of striatopallidal, striatonigral, and nigrostriatal projections in the Macaque. *J. Comp. Neurol* 304:569–595.

Hemmings, H. C., Walaas, S. I., Ouimet, C. C., and Greengard, P. (1987) Dopamine regulation of protein phosphorylation in the striatum: DARPP-32. *Trends Neurosci.* 10:377–383

Hoebel, B. G., Hernandez, L., Schwartz, D. H., Mark, G. P., and Hunter, G. A. (1989) Microdialysis studies of brain norepinephrine, serotonin and dopamine release during ingestive behavior: Theoretical and clinical implications. *Ann. N. Y. Acad. Sci.* 575:171–191.

Houk, J. C. (1992) Learning in modular networks. In K. S. Narendra (ed.), *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems.* New Haven, Conn.: Center for Systems Science, pp. 80–84.

Houk, J. C. (1992) *Learning in Modular Networks. NPB Technical Report 7,* Northwestern University University Medical School, Department of Physiology, Ward Building 5-342, 303 E. Chicago Ave., Chicago IL 60611-3008.

Houk, J. C., and Wise, S. P. (1993) Outline for a theory of motor behavior. In P. Rudomin, M. A. Arbib, and F. Cervantes-Perez (eds.), *Neuroscience: from Neural Networks to Artificial Intelligence, Research Notes in Neural Computing,* Vol. 4. Heidelberg: Springer-Verlag, pp. 452–470.

Houk, J. C., and Wise, S. P. (1994) Distributed modular architecture linking basal ganglia, cerebellum and cerebral cortex: Its role in planning and controlling action. *Cereb. Cortex* (submitted for publication)

Isaacson, J. S., Solis, J. M., and Nicoll, R. A. (1993) Local and diffuse synaptic actions of GABA in the hippocampus. *Neuron* 10:165–175.

Jiménez-Castellanos, J., and Graybiel, A. M. (1987.) Subdivisions of the dopamine-containing A8-A9-A10 complex identified by their differential mesostriatal innervation of striosomes and extrastriosomal matrix. *Neuroscience* 23:223–242.

Kawaguchi, Y., Wilson, C. J., and Emson, P. C. (1989) Intracellular recording of identified neostriatal patch and matrix spiny cells in a slice preparation preserving cortical inputs. *J. Neurophysiol* 62:1052–1068.

Kennedy, M. B., Bennett, M. K., and Erondu, N. E. (1983) Biochemical and immunochemical evidence that the "major" postsynaptic density protein is a subunit of a calmodulin-dependent protein kinase. *Proc. Natl. Acad. Sci. U. S. A.* 80:7357–7361.

Kita, H. (1992) Responses of globus pallidus neurons to cortical stimulation: intracellular study in the rat. *Brain Res.*

Kita, H., and Kitai, S. T. (1991) Intracellular study of rat globus pallidus neurons: Membrane properties and responses to neostriatal, subthalamic and nigral stimulation. *Brain Res.* 564:296–305.

Klopf, A. H. (1982) *The Hedonistic Neuron: A Theory of Memory, Learning and Intelligence.* New York: Hemispheres.

Lisman, J. (1989) A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proc. Natl. Acad. Sci. U. S. A.* 86:9574–9578.

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

Martinelli, G. P., Holstein, G. R., Pasik, P., and Cohen, B. (1992) Monoclonal antibodies for ultrastructural visualization of L-baclofen-sensitive GABA-B receptor sites. *Neuroscience* 46:23–33.

McGlade-McCulloh, E., Yamamoto, H., Tan, S.-E., Brickey, D. A., and Soderling, T. R. (1993) Phosphorylation and regulation of glutamate receptors by calcium/calmodulin-dependent protein kinase II. *Nature* 362:640–642.

Meyer, T., Hanson, P. I., Stryer, L., and Schulman, H. (1992) Calmodulin trapping by calcium-calmodulin–dependent protein kinase. *Science* 256:1199–1202

Minsky, M. L. (1963) Steps toward artificial intelligence. In E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought*. New York: McGraw-Hill, pp. 406–450.

Mody, I., Baimbridge, K. G., and Miller, J. J. (1984) Blockade of tetanic- and calcium-induced long-term potentiation in the hippocampal slice preparation by neuroleptics. *Neuropharmacology* 23:625–631.

Nairn, A. C., Hemmings, H. C., Jr., Walaas, S. I., and Greengard, P. (1988) DARPP-32 and phosphatase inhibitor-1, two structurally related inhibitors of protein phosphatase-1, are both present in striatonigral neurons. *J. Neurochem.* 50:257–262.

Newman-Gage, H., and Graybiel, A. M. (1988) Expression of calcium/calmodulin-dependent protein kinase in relation to dopamine islands and synaptic maturation in the cat striatum. *J. Neurosci.* 8:3360–3375.

Romo, R., and Schultz, W. (1990) Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *J. Neurophysiol.* 63:592–605.

Schulman, H., and Hanson, P. I. (1993) Multifunctional Ca^{2+}/calmodulin-dependent protein kinase. *Neurochem. Res.* 18:65–77.

Schultz, W., and Romo, R. (1992) Role of primate basal ganglia and frontal cortex in the internal generation of movements. I. Preparatory activity in the anterior striatum. *Exp. Brain Res.* 91:363–384.

Selemon, L. D., and Goldman-Rakic, P. S. (1990) Topographic intermingling of striatonigral and striatopallidal neurons in the Rhesus monkey. *J. Comp. Neurol.* 297:359–376.

Sutton, R. S. (1988) Learning to predict by the method of temporal differences. *Machine Learning* 3:9–44.

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

Sutton, R. S., and Barto, A. G. (1990) Time-derivative models of pavlovian reinforcement. In M. Gabriel and J. Moore (eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*. Cambridge, Mass.: MIT Press, pp. 497–537.

Weiler, I. J., and Greenough, W. T. (1993) Metabotropic glutamate receptors trigger postsynaptic protein synthesis. *Proc. Natl. Acad. Sci. U. S. A.* 90:7168–7171.

Werbos, P. J. (1992) Neurocontrol and supervised learning: An overview and evaluation. In D. White and D. Sofge (eds.), *Handbook of Intelligent Control*. New York: Van Nostrand Reinhold

Wickens, J. R. (1990) Striatal dopamine in motor activation and reward-mediated learning: Steps towards a unifying model. *J. Neural Transm.* 80:9–31.

Wilson, C. J. (1990) Basal Ganglia. In Shepherd, G. M. (ed.), *The Synaptic Organization of the Brain*. Oxford: Oxford University Press, pp. 279–316.

Wise, R. A., and Bozarth, M. A. (1984) Brain reward circuitry: Four circuit elements "wired" in apparent series. *Brain Res. Bull.* 297:265–273.

Wise, R. A., and Rompré, P.-P. (1989) Brain dopamine and reward. *Annu. Rev. Psychol.* 40:191–225.