Neuron, Vol. 36, 285–298, October 10, 2002, Copyright ©2002 by Cell Press

Reward, Motivation, and Reinforcement Learning

Review

Peter Dayan[1,3] and Bernard W. Balleine[2,3]
[1]Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
London WC1N 3AR
United Kingdom
[2]Department of Psychology and
the Brain Research Institute
University of California, Los Angeles
Box 951563
Los Angeles, California 90095

There is substantial evidence that dopamine is involved in reward learning and appetitive conditioning. However, the major reinforcement learning-based theoretical models of classical conditioning (crudely, prediction learning) are actually based on rules designed to explain instrumental conditioning (action learning). Extensive anatomical, pharmacological, and psychological data, particularly concerning the impact of motivational manipulations, show that these models are unreasonable. We review the data and consider the involvement of a rich collection of different neural systems in various aspects of these forms of conditioning. Dopamine plays a pivotal, but complicated, role.

Behavioral psychologists have long made a set of distinctions between classical/Pavlovian conditioning and instrumental/operant conditioning. Both forms of conditioning concern the ways that animals learn to predict and respond to important events in their environments, such as the delivery of appetitive and aversive stimuli (food/water and mild electric shocks, respectively). In classical conditioning, the outcomes are provided whatever the animals do, and so any changes in behavior presumably reflect innately specified reactions to predictions of the outcomes. Conversely, in instrumental conditioning, whether or not an animal gets a reward or punishment depends on the actions it performs. Instrumental conditioning is thus closely related to the engineering theory of optimal control and the computer science theory of reinforcement learning, which both study how systems of any sort can choose their actions to maximize rewards or minimize punishments. The operational distinction between classical and instrumental conditioning may seem like the sort of thing about which only dyed-in-the-wool behaviorists could care. However, critical issues turn on it, such as the organization of goal-directed behavior and motivational control, e.g., the way that animals head for water if thirsty but food if hungry. In this review, we argue that powerful and predictive theories of the neural basis of appetitive conditioning have failed to respect important aspects of the distinction and point the way toward their improvement.

The idea that dopamine cells in the vertebrate midbrain report errors in the prediction of reward has been a powerful (though not undisputed) organizing force for a wealth of experimental data (see Schultz et al., 1997; Schultz, 2002 [this issue of *Neuron*]; Montague and Berns, 2002 [this issue of *Neuron*]). This theory derives from reinforcement learning (Sutton and Barto, 1998), which shows how a particular form of the error (called the temporal difference error) can be used to learn predictions of reward delivery and also how the predictions can be used to learn to choose an adaptive course of action in terms of maximizing reward and minimizing punishment (Houk et al., 1995; Suri and Schultz, 1998, 1999). A popular reinforcement learning model called the actor-critic has been used to offer an account of the main subdivisions of the midbrain dopamine system. In one version of this, the dopamine cells in the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc) report the same prediction error, but turn it to two different uses. VTA dopamine cells are associated with the critic, controlling the learning of values held in the basolateral nucleus of the amygdala and the orbitofrontal cortex. SNc dopamine cells are associated with the actor, controlling the learning of actions in competitive cortico-striato-thalamo-cortical loops (Alexander and Crutcher, 1990).

Here, we evaluate the actor-critic model of the dopamine system from the perspective of the substantial psychological and neurobiological data on motivation, which is key to the modern view of reward learning. The actor-critic model can be seen as a particular form of one of the better accepted motivational theories, namely incentive learning theory (Bindra, 1974, 1978; Bolles, 1975; Toates 1986, 1994), in which stimuli associated with positive reinforcers such as food or water (which are known as Pavlovian excitors) are thought to act as conditioned incentives for instrumental conditioning. This is the conditioned reinforcement effect, that if a rat observes that a particular light always comes on just before it receives some food, it will then learn to press a lever in order to get that light, even if in those learning trials no food is provided. Despite this solid foundation, the actor-critic model fails to take into account large bodies of work (recently reviewed from a theoretical perspective in Berridge, 2001; Dickinson and Balleine, 2002) on the many psychological differences between Pavlovian and instrumental conditioning in terms of motivational processes and action determination and also on the neural underpinnings of these differences.

These data, together with those arguing against a simple conflation of Pavlovian conditioned responses, goal-directed instrumental actions, and habitual instrumental actions, force substantial changes to both aspects of the actor-critic model. Most critically, the foundations of the relationship between Pavlovian and instrumental conditioning in the existing dopaminergic model are decimated by recent data suggesting that there are at least two independent, largely anatomically distinct, reward processes, only one of which appears to be sensitive to dopaminergic manipulations. Modifications to the critic are consequent on evidence as to

[3]Correspondence: dayan@gatsby.ucl.ac.uk (P.D.), balleine@psych.ucla.edu (B.W.B.)

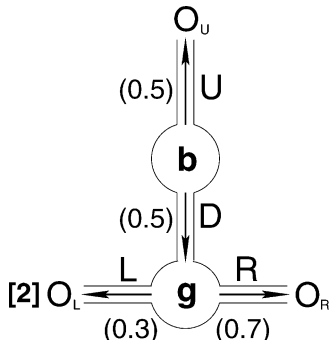

Figure 1. Example Task

Animals are presented with two stimuli (blue [**b**] and green [**g**] lights) defining two locations in a maze, at each of which they can choose one of two directions (U, D, R, and L). The actions have the consequences shown, leading to outcomes that are either neutral (O_U) or potentially appetitive (O_L and O_R). In brackets opposite the actions are the probabilities with which each is selected by a particular policy in a case in which food reward worth a nominal two units is available at O_L. This is a conceptual task; spatial mazes may engage different solution mechanisms.

how shifts in primary motivational states (such as hunger and thirst) exert an immediate effect on behavior. We suggest that there are two routes to the Pavlovian incentive predictions of the critic: a plastic one, copied from the original model, and a hard-wired one, with which it competes and whose output is directly sensitive to classes of motivational manipulations. Conditioned stimuli have access to this hard-wired route via stimulus substitution, another aspect of conditioning omitted from the original model.

The next section briefly describes the basic actor-critic. Following that, we review psychological and neural data on the Pavlovian control over actions and the differences between Pavlovian and instrumental motivational processes and use it to criticize the actor-critic model. Finally, we use it to sketch the bones of an alternative account, which is in better keeping with the data.

The Actor-Critic

Figure 1 describes an example task that we use to help organize the relevant data. It is not an exact replica of any real behavioral task. For convenience, think of it as a maze task for a rat, with two rooms, signaled by **b** (blue) or **g** (green) lights, out of each of which is two paths. At **b**, the subject can go "U" (up) or "D" (down); at **g**, it can go "L" (left) or "R" (right). Going U at **b** leads to an exit without any reward (outcome O_U); going L (left) or R (right) at **g** leads to exits with outcomes O_L (in some cases food, in others, nothing) or O_R (water or nothing). The task is clearly instrumentally posed in that the outcome for the subject is determined by its actions. The rat is presented with the task repeatedly, always starting at **b**, and can learn a good course of action over successive trials.

Consider the case in which the rat is hungry and food worth two notional units is available at O_L, whereas nothing is available at O_R or O_U. The rat has to learn to choose D at **b** and then L at **g** in order to get the food. This is an example of a sequential decision problem because the animal has first to choose the correct action at **b** (namely D) in order to be able to perform the correct action at **g** (namely L) in order to get any reward. Sequential decision problems (at least richer ones than this) are tricky to solve. For instance, imagine that the rat chooses to run D at **b** and then R at **g**. It gets no reward ($O_R = 0$); but, how can it decide whether it performed correctly at **b** and incorrectly at **g**, correctly at **g** but incorrectly at **b**, or indeed incorrectly at both? This ambiguity is one form of a problem called a temporal credit assignment problem. As we will see, in the actor-critic method, the critic learns to predict the expected reward starting from each room (**b** or **g**), and its predictions are used to resolve ambiguities.

Key to the actor-critic method (Barto et al., 1983) is a policy, which is a parameterized specification of how likely each action is to be chosen at each state (we use the notation $p_L(\mathbf{g})$ for the probability of choosing to go L at **g**). For instance, when the subject is put into the maze for the first time, a natural policy for it would specify that $p_U(\mathbf{b}) = 0.5$, $p_D(\mathbf{b}) = 0.5$, $p_L(\mathbf{g}) = 0.5$, $p_R(\mathbf{g}) = 0.5$, if it has no intrinsic bias at either location. The optimal policy, i.e., a policy which maximizes the reward the rat will get, has $p_U(\mathbf{b}) = 0$, $p_D(\mathbf{b}) = 1$, $p_L(\mathbf{g}) = 1$, $p_R(\mathbf{g}) = 0$. The action probabilities in brackets in Figure 1 indicate another policy. The actor specifies the policy.

The other half of the actor-critic method is the critic. This evaluates states under policies in terms of the average future reward starting from each. Take the policy shown in Figure 1. The value $V(\mathbf{g})$ of state **g** is defined naturally as the mean reward available in the task following the policy

$$\begin{aligned} V(\mathbf{g}) &= \text{probability of L} \times \text{value of reward given L} \\ &+ \text{probability of R} \times \text{value of reward given R} \\ &= 0.3 \times 2 + 0.7 \times 0 \\ &= 0.6. \end{aligned} \quad (1)$$

By contrast, under the optimal policy, $V(\mathbf{g}) = 2$ since the rat always collects the two units of reward if it starts at **g**. In conditioning terms, **g** acts as a conditioned stimulus (CS) with respect to the unconditioned stimulus (US; the food) provided as the outcome O_L. The value $V(\mathbf{b})$ of state **b** is also defined as the mean reward available in the task following the policy, but this time including the effect not only of the first action (U or D), but also the second action, only available here if D is selected. There are two ways to write this value, one directly:

$$\begin{aligned} V(\mathbf{b}) &= \text{probability of D} \times \text{probability of L} \\ &\times \text{value of reward} = 0.5 \times 0.3 \times 2 = 0.3, \end{aligned} \quad (2)$$

and one indirectly, using the fact that the only way to food is through state **g**:

$$\begin{aligned} V(\mathbf{b}) &= \text{probability of D} \times \text{value of state g} \\ &= 0.5 \times V(\mathbf{g}) = 0.3. \end{aligned} \quad (3)$$

Again, by contrast, under the optimal policy, $V(\mathbf{b}) = 2$ also, since the rat always collects two units of reward starting from **b**, by going through **g**. This points out an important characteristic of the critic, namely that its

evaluations are based on the sum total reward for the whole task, not just the immediate, next reward.

Equation 3 contains an example self-consistency condition between state values in that it specifies a quantitative relationship between the values V(**b**) and V(**g**). The temporal difference (TD) learning model of Pavlovian conditioning uses such relationships to allow the critic to solve its own temporal credit assignment problem. That is, the value V(**b**) of state **b** depends on distal rewards, i.e., rewards that the rat does not receive immediately following the action U or D that it performs at **b**. However, if it already knows V(**g**), then by measuring the inconsistency between its estimates of V(**b**) and V(**g**) when it runs from **b** to **g** on choosing action D, it can improve its estimate of V(**b**). The TD error, usually called δ, and mentioned above as the model for the phasic activation of dopamine cells, is just this inconsistency.

In the end, the point of learning the values is to improve the policy. That is, even though, for instance, neither U nor D produces any reward directly from **b**, D leads to state **g**, which is valuable, and therefore it is worth changing the policy to repeat D more frequently. Thus the critic trains the actor.

Recent neural implementations of TD models have proposed that the inconsistency between successive predictions, δ, is reported by a dopamine signal that controls synaptic plasticity (e.g., Montague et al., 1996; Schultz et al., 1997). A particularly compelling characteristic of the actor-critic model is that the very same (i.e., dopaminergic) signal, δ, can simultaneously be used to learn the values, putatively thought to involve ventral tegmental dopaminergic inputs to the amygdala, prefrontal cortex, and nucleus accumbens, and to improve the policy, with the latter potentially involving dopaminergic signals from substantia nigra to the striatum and elsewhere. Hence, the actor-critic model provides a powerful, unifying approach to both the reward and error correcting functions assigned to midbrain dopamine neurons.

Specifying the actor-critic model fully requires answering a number of key questions about the information associated with the predictions V(**b**) and V(**g**) and the policy. Various versions of these questions were first posed in the context of the associative-cybernetic model of Dickinson and Balleine (1993; see also Dickinson, 1994). For instance, is V(**g**) based on a direct association between stimulus **g** and average reward value (0.6 units), between stimulus and outcome (the food and/or its taste), or a more complicated representation that ties this prediction to the execution of action L? Similarly, is V(**b**) based on a direct association between stimulus **b** and average reward value (0.3 units), between stimulus **b** and stimulus **g** (for instance a predictive model of the task), or something involving actions D and/or L? Is the representation of the policy, p_L(**g**), based on a direct association between stimulus **g** and the action system or does it depend on V(**g**) too or a prediction that food will be delivered? These are critical questions because, for instance, the neural structures supporting stimulus-value predictions (e.g., **g**-0.6 units), appear to differ from those supporting stimulus-stimulus predictions (e.g., **g**-taste). Fortunately, psychological and neural data from sophisticated paradigms exist that help answer these questions. These paradigms probe what information is made available (for instance about the reward value of food to a sated rat) by which neural systems (notably dopamine, and parts of the nucleus accumbens) and at what points during the selection, or the learning of the selection, of actions in such tasks.

Actions, Habits, and Incentives

In the following sections, we consider two aspects of the data on reward learning that present a significant challenge to the actor-critic model, namely the Pavlovian specification of actions, and the differing motivational sensitivities of Pavlovian and instrumental incentive values. These falsify the simple actor-critic model and significantly constrain its successors. We will see that the structure of instrumental behavior changes from an early phase (during acquisition) to a later phase (following substantial or over-training), neither of which is accurately captured by the actor-critic model.

Pavlovian and Instrumental Actions

In the standard mapping of the actor-critic to conditioning, the critic, as a predictor of future reward and punishment, is thought to be a model for Pavlovian conditioning. However, real Pavlovian conditioning concerns more than just predictions, extending to the behavioral consequences of the predictions, namely conditioned responses (CRs). These CRs are reflex actions whose appropriateness is determined more by evolutionary processes than individual learning. For instance, take the case that turning on a light predicts the delivery of reward. The Pavlovian consequence of this is approach, i.e., the animal will move toward the light. This CR will be performed irrespective of whether or not it is appropriate in terms of gaining access to the reward, an outcome which is particularly striking when the experimenter arranges that food delivery depends upon withholding or reversing the direction of the CR. For example, Hershberger (1986) trained cochral chicks to expect to find food in a specific food cup. He then arranged the situation such that if they ran toward the food cup, the cup receded at twice their approach speed whereas if they ran away from the food cup, it approached them at twice their retreat speed. As such, the chicks had to learn to run away from the distinctive food cup in order to get food. Hershberger found that the chicks were unable to learn this response in order to get the food and persisted in chasing the food away. They could, however, learn perfectly well to get the food when the cup moved away from them at only half of their approach speed. Holland (1979) has conducted a conceptually similar experiment using hungry rats. He paired a light with food delivery, a preparation that quickly results in rats learning to approach the food source during the light. Holland compared performance in two groups, one for which approaching the food source during the light resulted in the omission of the food, and a second that received delivery of the food at the same time as the first group without regard to their responses. Even though the first group, the omission group, lost a considerable amount of food for doing so, Holland found that they acquired and maintained an approach response during the light to a similar degree as the second group for which there was no response contingency.

In terms of the actor-critic model, there is no justification for the performance of CRs at all, let alone ones that act to reduce the delivery of reward. By itself, this is an important lacuna of the model. Various facets of CRs are known. For instance, it has long been recognized that Pavlovian CRs differ from responses generated by the US (cf. Wagner and Brandon, 1989, for discussion). Nevertheless, they are generally regarded as reflexive in nature, reflecting both the sensory properties of the CSs (like tones and lights) (Holland, 1977) and the sensory and emotional or motivational properties of USs (like food and water) (Konorski, 1967). At a neural level, therefore, it appears likely that CRs will have much in common with simple stimulus-response (i.e., S-R) habits. Although there is little data to indicate how such Pavlovian habits are encoded, it is likely that the dorsal striatum and striatal-like regions of the amygdala and cerebellum along with their connections with midbrain and brain stem motor nuclei contribute to the performance of these simple responses (Holland, 1993; Gluck et al., 2001).

Although Pavlovian actions can be highly adaptive (e.g., Hollis et al., 1997), the fact their performance can be less than appropriate in a changing environment makes it necessary that animals be able to acquire new behavioral strategies in order to get essential commodities. This well-documented capacity is usually studied within the free-operant or instrumental conditioning procedure; in the paradigm case, a hungry rat can be trained to press a freely available lever to gain access to food reward. The key questions for us are exactly what is learned about the action and the outcome (e.g., the food), and whether, as in the actor-critic, this learning depends at all on Pavlovian predictions? Considerable evidence suggests that at least during acquisition, the performance of instrumental actions depends crucially on encoding the relationship between the action and its consequences or outcome (i.e., the instrumental contingency; see Dickinson and Balleine, 1993, 1994; Colwill and Rescorla, 1986, for reviews). The extent to which this dependence is true after, as well as during, learning bears importantly on the way that policies such as $p_L(\mathbf{g})$ are represented and is a major topic in the section of this paper on Instrumental Values. Note, however, that in any given situation, it is not necessarily immediately apparent whether the performance of a candidate action is controlled by an instrumental or a Pavlovian contingency, and tests have to be conducted to establish this. To illustrate this point, consider the example presented in Figure 1. A hypothetical rat might be observed to perform D more than U in the presence of the state cue **b**. But this could be due either to the relationship between D and **g** (the instrumental contingency) or to the relationship between **b** and **g** such that **b** elicits D independently of D's relationship to **g** (the Pavlovian contingency), just like the chicks' prediction of food produces an approach response independently of the consequence of that response for the delivery of the food. Although in this situation, it makes little difference with respect to the transition to **g** whether a Pavlovian or an instrumental contingency controls the performance of D, it is clear that it will have direct implications for anticipating the neural structures that control D.

In instrumental conditioning, when the relationship between the performance of an action (such as pressing a lever) and its consequences (such as food) is changed, the performance of the action changes appropriately. Hungry rats will, for example, reduce their performance of actions as the contingent relation between the action, and the delivery of a specific food outcome is reduced (Hammond, 1980; Dickinson and Mulatero, 1989; Balleine and Dickinson, 1998). Indeed, they will reduce their performance even more rapidly if a negative relationship is arranged, e.g., if performing the action leads to the omission of an otherwise available food reward (Davis and Bitterman, 1971; Dickinson et al., 1998). Omission schedules provide, therefore, a direct means of assessing the relative degree of Pavlovian and instrumental control over behavior (cf. Dickinson and Balleine, 1993).

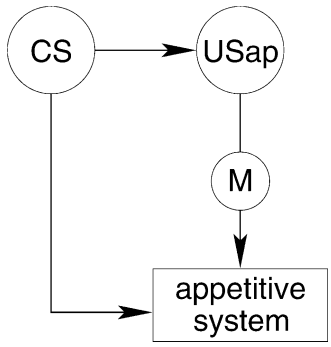

Figure 2. Model of the Pavlovian Motivational System

An appetitive US representation is connected with the appetitive affective system via a motivational gate (M) that is sensitive to the specific biological impact of the US and that maintains a fixed connection with the appetitive affective system. CSs can form connections with the US or the appetitive system directly. Modeled after Dickinson and Balleine, 2002.

Interestingly, when instrumental actions are overtrained, they no longer appear to adjust to the imposition of an omission contingency (Dickinson et al., 1998). This has been taken as evidence that a simple sensory-motor or S-R habit has been formed, which then acts in essentially the same manner as a Pavlovian conditioned response. This shift in the associative structure controlling instrumental performance provides some very strong constraints on models by showing how the information determining policies such as $p_L(\mathbf{g})$ changes over time.

Pavlovian Values

Most dopaminergic models of the critic assume that Pavlovian CRs, such as approach to a food source, offer a behavioral report of Pavlovian values like $V(\mathbf{b})$ and $V(\mathbf{g})$. Although Pavlovian CRs are not controlled by the CR-US contingency, their performance is directly sensitive to the level of motivation of the subjects, such as whether or not they are hungry. This motivational or affective control argues against simple realizations of the critic in which there is one set of weights or parameters (whatever their neural realization) mapping from stimulus representations (i.e., units whose activities are determined by the stimuli shown to the subject) to the values. Rather, it supports something more like the fundamentally richer model of Figure 2 (Dickinson and Balleine, 2002).

Dickinson and Balleine (2002) argue that there are two

different kinds of connections between stimuli and the appetitive and aversive affective structures that control performance of CRs: one direct and one indirect via a connection between the CS and sensory properties of the US. Evidence for the direct connection between CSs, and an appetitive system is provided by demonstrations of transreinforcer blocking. That is, in rats made both hungry and thirsty, a CS paired with food can subsequently block the acquisition of conditioned responses to a second CS when both CSs are presented in compound and reinforced with water (Ganesan and Pearce, 198). A standard interpretation of blocking is that the first stimulus makes an accurate prediction, leaving nothing for the second stimulus to predict. For this to hold when the reinforcer is changed across phases, however, it must be true that the affective value of a CS can be divorced from the particular US that originally established that value, as in the direct pathway. There is, however, equally good evidence for the indirect connection, that is, stimulus-stimulus (stimulus substitution) links between CSs and USs that allow CSs to activate the appetitive system transitively. Furthermore, the classic reports of the sensitivity of CRs to shifts in motivation (DeBold et al., 1965; Mitchell and Gormezano, 1970; reviewed by Dickinson and Balleine, 2002) suggest that the impact of this CS-US association on performance can depend on the motivational state of the subject. As a consequence, motivational states such as hunger and thirst have been argued to threshold or gate (i.e., M in Figure 2) connections between the sensory US representation and the appetitive system and, therefore, to modulate the indirect link from CSs to the appetitive system.

In this model, the gate is US specific, so that, for instance, it is hydration that modulates fluid representations and nutritional needs that modulate food representations. Indeed, studies of specific hungers suggest that US gating may be even more specific than a simple hunger-thirst distinction. For example, Davidson et al. (1997) exposed food-deprived rats to pairings of one CS with a carbohydrate US, sucrose pellets, and another with a fat US, peanut oil, before returning the animals to the nondeprived state for testing. Prior to this test, one group was placed in a glucoprivic state (i.e., short of sugar), whereas a lipoprivic state (i.e., short of fat) was induced in another group. Importantly, the rats in the lipoprivic state showed more magazine approach during the peanut oil CS than during the sucrose CS, whereas the reverse pattern was observed in rats in the glucoprivic state. This specificity implies that there must be separate US gating for these two states and their relevant reinforcers. Another example of this specificity is salt seeking under a sodium appetite (e.g., Berridge and Schulkin, 1989; Fudim, 1978; see also Krieckhaus, 1970).

We can now see how the values assigned to state cues in Figure 1 should be affected by this Pavlovian incentive process. To the extent that the indirect **g**-US pathway of Figure 2 applies, the value V(**g**) of state **g** will be dependent on the motivational control of the **g**-O_L association and so will be high, provided the hypothetical rat is hungry, and low when it is satiated. The value V(**b**) of state **b** depends partly on a second order contingency, since **b** is followed by **g** rather than directly by reward. Under certain conditions (e.g., Rescorla, 1982), responding to second-order cues has been reported to be insensitive to the effects of a shift in motivation. Under these conditions, V(**b**) may be maintained by the direct pathway to the affective system shown in Figure 2.

Interestingly, representations of sensory events in the primate orbitofrontal cortex have been argued to undergo remodeling based on their motivational significance; the responsiveness of neurons in this region sensitive to food or water delivery, as well as to signals that predict the delivery of these commodities has been found to be a direct function of the animal's motivational state (e.g., of their degree of food or water deprivation; cf. Rolls, 1989, 2000a). It appears that in other structures involved in ascribing affective significance to CSs, such as the insular cortex and the amygdala, neuronal responsiveness is not gated by motivational state (Yaxley et al., 1988; Rolls et al., 1990; Rolls, 2000b), suggesting that these areas may be involved in the direct pathway associating CSs to the appetitive system.

Since the tasks in which the activity of dopamine cells has been recorded (Schultz, 1998) have not involved motivational manipulations, we do not know the behavior of the prediction error signal for Pavlovian rewards in these cases. It has been suggested that dopamine is preferentially involved in the direct pathway of Figure 2. For instance, Berridge and his colleagues (see Berridge, 2001) have implicated dopamine in the preparatory (wanting) aspects of appetitive conditioning and have shown its irrelevance for, at least, those consummatory appetitive orofacial reactions elicited by the direct infusion of foods and fluids. Berridge (2001) and Balleine and Dickinson (1998) have specifically identified the gustatory region of the insular cortex and its efferents to the amygdala as mediating important aspects of this latter (i.e., liking) aspect of food rewards based on the finding that damage to this region appears to generate a form of taste agnosia (e.g., Braun, 1990).

Instrumental Values

Having established an observational difference between Pavlovian and instrumental actions and a sensitive mechanism for assessing Pavlovian values, we might seem to be in a good position to assess the key claim of the actor-critic model by asking how well matched to the data is a learning scheme that uses a dopaminergic Pavlovian critic to train an instrumental actor. Unfortunately, this turns out to be a rather complicated question. Broadly speaking, the model does not fare well except potentially in the case of conditioned reinforcement, for which some critical experiments have yet to be performed.

First, consistent with the actor-critic, signals (like **g** in Figure 1) associated with reward can act as conditioned reinforcers of actions. It has long been known that rats and other animals will acquire actions like lever pressing when those actions are followed by a cue previously associated with primary reward (e.g., food; see Mackintosh, 1974, for review). Furthermore, it is generally accepted that although the acquisition of responses for such conditioned reinforcers depends upon the basolateral amygdala, the reinforcement signal appears to rely on a Pavlovian incentive process involving dopaminergic activity in the ventral striatum (Cador et al., 1989; Taylor and Robbins, 1984). Unfortunately, because of the control conditions conventionally employed, it is not

yet known whether these actions are under Pavlovian or instrumental control. It is common, for example, to compare responding on a lever that delivers a stimulus paired with reward to performance on another lever that delivers a stimulus that has not been paired with reward. However, since the levers are usually spatially separated, a Pavlovian approach to the location of the lever delivering the paired stimulus would be as effective in causing that lever to be favored as choice based on the encoded action-outcome contingency.

Second, consider the representation of a policy such as $p_L(\mathbf{g})$. A standard simple actor scheme treats this as a learned stimulus-response mapping from **g** to an action selection system (often considered to involve the dorsolateral striatum and, particularly, on its large matrix compartment to which most sensory and motor fibers project; e.g., Gerfen, 1985; Graybiel et al., 1994). This is too simplistic, both during the acquisition and after the overtraining of an instrumental action. The problem for any theory of instrumental acquisition formulated in terms of S-R association is that if, for instance, the food reward is devalued by independently conditioning a taste aversion to it, then evidence suggests that the rat will refuse to go L at **g** when returned to the task. This implies that $p_L(\mathbf{g})$ depends directly on the value of outcome that is encoded as a consequence of the action and not on the relationship between stimulus **g** and response L. Indeed, this dependence is often taken as the mark of a truly instrumental action (e.g., Dickinson and Balleine, 1993, 1995). In modeling terms, it suggests that a forward model (e.g., Sutton and Pinette, 1985; Dayan, 1993; Wolpert and Ghahramani, 2000), which predicts the consequences of actions at states, must play a role in specifying the policies. One might hope that the simple actor would be better able to account for the end point of instrumental conditioning, by which time, as we have discussed, actions appear to become stimulus-response habits dependent on the stimulus **g** and the motivational state of the animal (Dickinson et al.,1995). However, because it fails to accommodate Pavlovian actions, the standard actor-critic lacks a way for Pavlovian values to control the execution of habits (but see the discussion of Pavlovian-instrumental transfer below) and would, contrary to the data (Dickinson et al., 1998), suggest that such habits would be readily unlearned through omission schedules.

Third and most critical for the actor-critic scheme of using Pavlovian values to control instrumental learning, is the considerable evidence suggesting that the instrumental incentive process conforms to different rules than Pavlovian incentives and has a different neural basis (Balleine, 2001; Corbit et al., 2001). Consider the hypothetical rat in Figure 1 in a slightly different case in which food is available at O_L and water at O_R . We know from the discussion in the Pavlovian Values section that the Pavlovian value of **g** has a gated dependence on motivational state and so will change appropriately as the animal is sated for food, water, or both. However, the same turns out not be true for truly instrumental actions. Dickinson and Dawson (1988) trained rats to lever press and chain pull with one action, earning access to liquid sucrose with one and the other to dry food pellets. When subsequently made thirsty, the rats did not immediately modify their performance on the levers and chains but continued to produce both responses at an equivalent rate. If, however, the rats were allowed merely to consume the liquid sucrose and dry pellet rewards when thirsty prior to the choice test, they subsequently displayed a strong preference for the action that, in training, had previously delivered the liquid sucrose. Although Pavlovian responses have been found to be directly sensitive to this shift (e.g., Dickinson and Dawson, 1987), the shift to water deprivation had no impact on instrumental performance until the rats had been allowed to experience the effect of the shift on the incentive value of the instrumental outcome. The latter training, involving consummatory contact with the instrumental outcome after a shift in primary motivation, is called incentive learning (see Dickinson and Balleine, 1994; Balleine, 2001, for reviews).

The two parts of this finding jointly pose a complicated problem for actor-critic models. The simple actor-critic is happily insensitive to changes in motivational state. That incentive learning about the outcomes can affect a subsequent choice of action suggests the involvement of a forward model to allow information about an outcome to control a determining action via a prediction. Indeed, Balleine (2001) argues in general for this sort of indirect connection between instrumental actions and motivational structures as the basis for the primary representation of the hedonic properties of the instrumental outcome. However, the evaluation of the outcome cannot be the Pavlovian evaluation that the actor-critic would have assumed since Pavlovian values do not require incentive learning to come under motivational control (cf. Balleine, 1992, 2001).

Incentive learning is very general. For example, rats trained to lever press for food when food deprived do not immediately reduce their performance on the lever when they are suddenly shifted to an undeprived state. Nor do they increase their performance immediately if they are trained undeprived and are suddenly given a test on the levers when food deprived. In both cases, rats only modify their instrumental performance after they have been allowed the opportunity to consume the instrumental outcome in the new motivational state (Balleine, 1992). Nor is this effect confined to shifts between hunger and thirst and hunger and satiety. It has also been observed after increases and decreases in water deprivation (Lopez et al., 1992; Lopez and Paredes-Olay, 1999), devaluation by taste aversion procedures (Balleine and Dickinson, 1991, 1992), changes in outcome value mediated by drug states (Balleine et al., 1994), and even following changes in the value of thermoregulatory (Hendersen and Graham, 1979) and sexual rewards (Everitt and Stacey, 1987). Nevertheless, the neural processes that support incentive learning are not well understood at present, although Balleine and Dickinson (2000) have presented evidence to suggest that in instrumental conditioning, the insular cortex is involved in the retrieval of the incentive value of foods. At a systems level, afferents from the insular cortex to the amygdala and orbitofrontal cortex, together with connections between these latter structures, may provide the essential circuit mediating learning about changes in the rewarding impact of instrumental outcomes following shifts in primary motivation. To the extent that that is true, connections between these struc-

tures and cortico-striatal circuits involving the prelimbic area, the nucleus accumbens core, and pallidal output from the basal ganglia may provide the means by which these changes in reward act to affect instrumental performance (cf. Balleine and Dickinson, 1998, for discussion).

We have thus argued that instrumental and Pavlovian values are separately represented and that dopamine plays an important role in the direct ascription of Pavlovian values to CSs. We also know that dopamine plays an important role in conditioned reinforcement, although it is not yet clear whether or not conditioned reinforcement involves instrumental conditioning. Strikingly, dopamine seems not to play a role in representing the sort of instrumental values involved in incentive learning. However, this independence is confusingly masked by a dopamine-dependent interaction between Pavlovian and instrumental conditioning called Pavlovian-instrumental transfer.

It is well documented that Pavlovian cues can exert an excitatory effect on instrumental performance. In humans, for example, cues associated with drugs of abuse, in addition to inducing a strong craving for the specific drug with which they are associated, produce a rapid and selective increase in actions through which addicts seek access to the drug and, even after long periods of abstinence, reliably precipitate relapse to a drug-taking habit (O'Brien et al., 1998) In rats, this phenomenon has been studied using several different approaches. For example, consider training a rat separately that a tone CS is associated with water, whereas pressing a lever leads to food. If the rat is subsequently made thirsty, then playing the tone while it is pressing the lever results in an increase in performance on the lever. This is the basic Pavlovian-instrumental transfer effect (cf. Balleine, 1994). More impressive still is evidence for selective effects of Pavlovian CSs; for example, Colwill and Rescorla (1988) and Colwill and Motzkin (1994) have reported that CSs that signal the same reinforcer as that earned by an instrumental action (i.e., if the tone was associated with food rather than water) facilitate the performance of that action but have no effect on the performance of actions trained with a different reinforcer to that signaled by the CS.

Considerable evidence suggests that like conditioned reinforcement, Pavlovian-instrumental transfer involves the ventral striatum (de Borchgrave et al., 2002; Balleine and Killcross, 1994) and, more specifically, dopaminergic activity in the nucleus accumbens (e.g., Wyvell and Berridge, 2000, 2001; Dickinson et al., 2000). For instance, the impact of Pavlovian cues on instrumental performance is strongly affected by negative effects on striatal DA (such as lesions) (Berridge and Robinson, 1998), or the administration of pimozide (Pecina et al., 1997), or the facilitation of DA transmission by either microinjection of amphetamine into the shell region of the nucleus accumbens (Wyvell and Berridge, 2000), or amphetamine-induced sensitization (Wyvell and Berridge, 2001).

Critically, two lines of evidence argue that this pathway is not associated with instrumental values. First, treatments that modify the effectiveness of Pavlovian cues on instrumental performance have no detectable effect on the sensitivity of instrumental performance to devaluation of the instrumental outcome. In one study, for example, peripheral administration of either the D2 antagonist pimozide or the D1, D2 antagonist α-flupenthixol were found to induce both a dose-dependent decrease in instrumental lever pressing for food and to attenuate the excitatory effects of a Pavlovian CS for food on instrumental performance. Nevertheless, neither drug was found to influence the instrumental devaluation effect induced by a shift from a food deprived to a nondeprived state (Dickinson et al., 2000). Dickinson et al. concluded that the changes in the incentive value of the instrumental outcome induced by devaluation treatments are mediated by a different process to that engaged by excitatory Pavlovian cues; whereas the latter appears to be dopamine dependent, the former does not.

The potentiation of Pavlovian-instrumental transfer induced by infusions of amphetamine into the accumbens shell suggests that this region of the accumbens may critically mediate the excitatory effects of Pavlovian cues on instrumental performance without affecting outcome devaluation. Direct evidence for this claim comes from a recent series of studies by Corbit et al. (2001). In this series, selective lesions of the accumbens shell were found to abolish the selective transfer effects produced when a CS is paired with the same reinforcer as that earned by the instrumental action. Nevertheless, no effect of this lesion was found on the sensitivity of rats to selective devaluation of the instrumental outcome by a specific satiety treatment. Corbit et al. (2001) compared the impact of shell lesions with lesions made of the accumbens core. Importantly, lesions of the core were found to have no influence on the selective transfer effect abolished by the shell lesions but had a profound effect on the sensitivity of rats to the selective devaluation of the instrumental outcome. Corbit et al.'s (2001) study provides evidence, therefore, of a double dissociation between the impact of shell and core lesions on Pavlovian transfer and instrumental devaluation effects, suggesting that Pavlovian and instrumental incentive processes involve distinct neural systems.

The New Model

Before outlining the new model designed to accommodate these findings, it is worth noting again how unexpected they are from the perspective of methods for learning optimal actions. The Pavlovian system is associated with an extremely rigid (and, indeed, often maladaptive) scheme for action choice. Stimuli associated with appetitive USs elicit approach-like behavior (as in the pecking of the food- or water-predicting lit key in pigeon autoshaping) whether or not (as in omission schedules) this behavior is adaptive or even barely motivationally appropriate. However, coupled with this system for choosing actions is a highly sophisticated method for evaluating the motivational relevance of USs and their predictors, which instantly reevaluates them in the light of the animal's multidimensional needs for water, food, salt, lipids, and the like.

The instrumental system allows much greater flexibility in the choice of actions, permitting the acquisition of relatively arbitrary chains of responses. However, it suffers from an incentive system that seems to lack the

capacity for instant reevaluation of rewarding events in the light of shifts in the current motivational state. Rather, motivational states seem to act merely as part of the representational context for specifying values, and only through learning (notably incentive learning) can motivational state be tied to instrumental values. Of course, this opportunity to establish values by learning can be seen as another way in which the instrumental system is more flexible than the Pavlovian system in its ability to adapt to changes in environmental constraints. Nevertheless, a Pavlovian appetitive conditioned stimulus (e.g., a light that predicts food to a hungry animal) can act as conditioned reinforcer, although, in this situation, it is possible that responding is controlled by the Pavlovian, rather than the instrumental, contingency.

In this section, we sketch the outline of a new account (Dayan, 2002) designed in the light of these data and Dickinson and Balleine's (1993) associative-cybernetic model. It relates Dickinson and Balleine's (2002) Konorskian model of Pavlovian motivation shown in Figure 2 to the dopamine data and formalizes a model of instrumental action choice that can be related to incentive learning. The most complex aspects of the model come from the existence of two different predictive models: one for Pavlovian conditioning seen in the CS-US association in Figure 2, and one for instrumental conditioning. There is also a complex interplay between hard-wired values (as in the motivational gate M in Figure 2) and learned or plastic values.

Formally, the new model, like the actor-critic, considers predictions [such as V(**b**)] of the long-run future rewards starting from each state and following a policy. Unlike the actor-critic, it evaluates actions according to quantities called their advantages (Baird, 1993). The advantage of performing an action at a state is the difference between two long-term rewards: one particular to the action, called its Q value (Watkins, 1989), the other averaging over all actions (the regular value of the state). For instance, the advantage of action D at state **b** is:

$$A_D(\mathbf{b}) = Q_D(\mathbf{b}) - V(\mathbf{b}), \quad (4)$$

where $Q_D(\mathbf{b})$ is the expected reward if the rat goes D at **b** and then follows its normal policy from the state at which it arrives, and V(**b**) is the regular prediction of long run future reward from **b**. For the policy shown in Figure 1, $Q_D(\mathbf{b}) = 0.6$ is just the value V(**g**), since choosing D at state **b** leads the rat to state **g**. This implies that the advantage of action D at **b** is $A_D(\mathbf{b}) = 0.6 - 0.3 = 0.3$. Conversely, the advantage of action U at **b** is $A_U(\mathbf{b}) = 0 - 0.3 = -0.3$, since $Q_U(\mathbf{b}) = 0$. The advantages are so named because they measure how much better [$A_D(\mathbf{b}) > 0$] or worse [$A_U(\mathbf{b}) < 0$] an action is than the state's value, V(**b**), which comes from averaging across all actions at **b**.

Actions with large, positive advantages are better than the actions specified by the current policy and should be favored. Actions with large negative advantages are poor and should be disdained. For the policy in Figure 1, this suggests that the animal would do better by choosing action D at **b** more frequently than action U, which is clearly true. This process by which action choice gets better is a form of policy improvement. In general, we assume that the rat learns the advantages of the actions and when faced with a choice, that the actions compete according to their advantages. The neural substrate for this competition is not clear, although cortico-dorsal striato-thalamo-cortical loops have been argued to play an important role (Reynolds et al., 2001).

As the choice of actions at a state changes, the value of that state and the advantages of all the actions at that state change too. At the optimal policy for the maze in Figure 1 for which $p_U(\mathbf{b}) = 0$, $p_D(\mathbf{b}) = 1$, $p_L(\mathbf{g}) = 1$, $p_R(\mathbf{g}) = 0$, and the Q value $Q_D(\mathbf{b}) = 2$, since the rat always chooses correctly at **g**, and so the advantage $A_D(\mathbf{b}) = 2 - 2 = 0$. Correctly (and importantly) the optimal action D has no advantage over the average value of all actions when it is itself always chosen at **b**.

In the model, values are learned using the familiar, phasic prediction error signal δ reported by dopamine (Montague et al., 1996; Schultz et al., 1997; Schultz, 1998). That is, the model for learning V(**b**) and V(**g**) is exactly as before. The value of this same prediction error signal following execution of an action is actually the target for the advantage of that action. Thus, the error in the advantage $A_D(\mathbf{b})$ is

$$\delta_A = \delta - A_D(\mathbf{b}), \quad (5)$$

based on the TD prediction error δ that arises after action D is taken at state **b**. The advantage error signal δ_A can be used to alter synaptic weights, which are responsible for $A_D(\mathbf{b})$.

We have so far described the key components of the model. The next task is to map them onto the neural and psychological data described above. The model adopts Dickinson and Balleine's (2002) scheme for Pavlovian evaluation (as shown in Figure 2), suggesting that a key output from the appetitive system is the dopaminergic temporal difference prediction error δ, which exerts Pavlovian motivational control over actions as well as training the values. This captures the neurophysiological data on the activity of the dopamine system (which have hitherto not involved motivational manipulations) and is also consistent with the other motivational effects discussed in the previous section. The model assumes that Pavlovian-instrumental transfer is based on the dopaminergic output of this system (likely to the shell of the accumbens) and that the same signal may control the vigor with which Pavlovian conditioned responses are emitted. The nature of the direct associative link between CSs and USs and also between CSs and other CSs, as in other Pavlovian paradigms such as sensory preconditioning (Suri and Schultz, 2001), is not completely specified. Certainly some form of stimulus-substitution is likely to play an important role. As suggested by Figure 2, the motivational predictions of CSs (the CS-appetitive pathway) are in a form of learning competition with the predictions based on stimulus substitution (CS-USap-M-appetitive). The larger the contribution of the latter to the final appetitive prediction of the CS, the more the CS will be subject to instant motivational manipulations, for instance, allowing a light that predicts that food will immediately follow might lose its apparent conditioning power if the animal is sated.

This dopaminergic prediction error signal might control the energy level applied to a Pavlovian habit but to

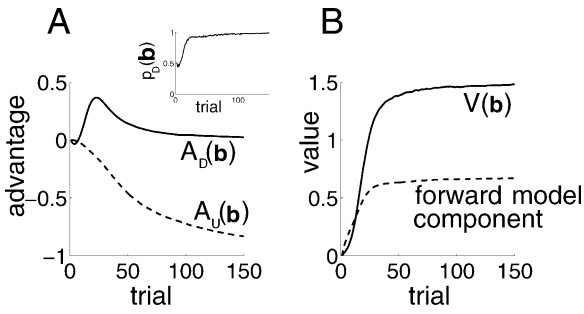

Figure 3. The New Model

(A) Development across trials of the advantages of D and U at state **b** in a case in which the animal has already been shaped to go L at state **g** and D costs a nominal −0.5 units. The inset graph shows the probability of choosing to go D at **b**.

(B) The development of the value V(**b**) of state **b** (solid line), which assumes responsibility for the control of going D once the action has become a habit (i.e., once the associated advantage is 0). The dashed line shows the component of the value that comes from a predictive or forward model.

which habit should it be applied? Here, as in Dickinson and Balleine (2002), we make another assumption: that USs such as food and water are directly associated with appropriate consummatory habits such as eating and drinking and that pure motivational predictions, i.e., predictions of future reward or future punishment, are associated with hard-wired preparatory habits such as approach or withdrawal. These habits extend by stimulus substitution to CSs such as lights and tones with which the USs are associated, albeit with some changes associated both with the different sensory impact of the CSs and their relative abilities to activate the US representations.

Next, the instrumental action selection mechanism, which allows learning of arbitrary action sequences for future reward, must provide a means by which stimulus-response habits can be born from actions. In the example of Figure 1, for instance, an action selection system in which the choice must be made between simultaneously appropriate possibilities U and D at **b**, under instrumental motivational control, must, over the course of learning, give rise to a habit scheme in which a single course of action (D) is appropriate. In the habit scheme, evidence suggests that the only control that remains is the degree of Pavlovian motivation [V(**b**)] that it enjoys (cf. Dickinson et al., 1995).

The new model captures this using the advantages. As discussed, over the course of policy improvement, the advantage of a suboptimal action (e.g., U) becomes negative [$A_U(\mathbf{b}) \rightarrow -2$] and of an optimal action (D) tends to 0. The latter offers a natural model of the transition from instrumental action selection to a stimulus-response habit. Once the advantage of the optimal action has become 0, the only remaining control at a state becomes the Pavlovian prediction [V(**b**)] associated with that state, i.e., the Pavlovian motivational system takes over.

Figures 3A and 3B show two aspects of instrumental conditioning at state **b** in the task of Figure 1. Here, we have assumed that the subject has already been shaped to choose L at **g**, but, to show off the full characteristics of the scheme, make the action D have a small cost (of −0.5 units). Figure 3A shows the development of the advantages of U and D over learning. Action D starts looking worse because it has a greater immediate cost; its advantage increases as the worth of D grows greater than the mean value of **b** and then goes to 0 (the birth of the habit) as the subject learns to choose it every time. The small inset graph shows how the probability of choosing D grows toward one over trials. For comparison, the solid line in Figure 3B shows the value component V(**b**) of state **b**. While actions are still under instrumental control [here, up until around trial 100, when $A_D(\mathbf{b}) \rightarrow 0$], they can compete according to their advantages. One natural assumption is that there is a baseline for the advantages, and as the advantage of an action dips below that baseline, that action is removed from the competition. The dashed line in Figure 3B (labeled forward model) shows the component of V(**b**) that derives from the CS-USap-M-appetitive connection of Figure 2; the remainder depends directly on plastic connections between the CS and the appetitive system. This former component is subject to instant devaluation as in the motivational experiments.

The incentive learning experiments show that a predictive model plays a critical role in instrumental evaluation. Various experiments suggest that the cortical areas associated with taste processing are substrates for incentive learning. One way to think of these is rather like the hard-wired system for Pavlovian motivation. Actions predicted to lead to good-tasting substances are automatically awarded high advantages; those leading to bad-tasting ones are automatically awarded low advantages. If the taste changes through learning (for instance in the reexposure condition of a food aversion experiment), then this has an immediate impact on the advantage of the associated action. As with the Pavlovian system, the actual advantages result from the combination of the output of this route with that of a conventionally plastic route. From the data at the end of the section entitled Instrumental Values, the core of the accumbens may be an important player in determining the interaction of instrumental values with the instrumental action-outcome contingency (Corbit et al., 2001).

Unfortunately, because of the tasks, the data do not settle the issue as to exactly what the animals are predicting and under what conditions. In fact, there is an extreme model of instrumental action choice that is purely cognitive (Dickinson, 1997), using only the predictions of actual outcomes based on a full model of the action-state-outcome contingencies of the world. Since schedules involving many sequential choices suffer from curses of dimensionality, i.e., an explosion in the size of the relationship between action and outcome (think of the complexity of trying to look multiple moves ahead in chess, for instance), this extreme method of action choice is only effective in very small domains. Dynamic programming avoids this curse of dimensionality by using value functions as intermediate quantities. These functions, at the other extreme, eliminate the complexity in action choice by focusing exclusively on motivational value, divorced from its basis in the set of choices between actions. There is a variety of possibilities between these two extremes in which world models play a greater or lesser role. The theoretical reinforce-

ment learning community has yet to resolve the issues as to the regimes in which different forms of world model are most obviously necessary (Kearns and Singh, 1999).

As mentioned, the neural substrate of the instrumental actor is not completely clear. There are general reasons to suppose that the dopaminergic pathway from the substantia nigra pars compacta (SNc) to the dorsal striatum might report the error in the current advantage error of an action at a state, and cortico-striato-thalamo-cortical loops might be responsible for action competition. However, there is not yet the evidence to pin down even whether the substrate for choosing a response changes as the response goes from being under instrumental to Pavlovian control.

Discussion

Experiments pose a critical challenge to our understanding of the psychological and neural implementation of reinforcement learning, suggesting the importance of two different sorts of motivation in controlling behavior and arguing against the simple actor-critic scheme that has hitherto generally been assumed. The experiments also place significant emphasis on habits, the objects of Pavlovian motivation, which are orphaned in the previous scheme.

We adopted Dickinson and Balleine's (2002) characterization of Pavlovian motivation, incorporating a hardwired, stimulus substitution-sensitive route for the evaluation of stimuli and states, which competes with a plastic route putatively operating through the amygdala and the orbitofrontal cortex. One key output of this model, reported by dopaminergic activity, is the temporal difference prediction error for future reward. This, acting via the shell (and possibly also the core) of the nucleus accumbens, is suggested as the mechanism underlying the Pavlovian control of habits.

We also considered advantages as a suitable way of modeling the transition from the selection between multiple, simultaneously appropriate actions under instrumental control, to the Pavlovian-based motivation of just a single action. One component of the advantage must be dependent on a forward or predictive model, so that instrumental action choice can be sensitive to reexposure as in incentive learning.

A more subtle message of these studies is the complexity of the learning phenomena and the neural substrates involved. Sequential decision problems involving chains of actions, as in Figure 1, are highly revealing because motivational and other manipulations readily lay bare the extent to which actions are chosen and values are specified based on prewired direct and indirect (i.e., forward-model) links between CSs and USs and between actions and their associated discriminative CSs and outcome USs. The differing sensitivities discovered under the heading of incentive learning show the variegated roles played by different parts of the brain involved in appetitive evaluation.

The main theoretical problem left concerns exploratory behavior and the trade-off between exploring to gain new information about an environment and exploiting existing knowledge in order to get reward. Although formal solutions to this problem are well known, they are highly computationally challenging. Studying how animals choose (and learn to choose, Krebs et al., 1978) to balance exploration and exploitation is a rich field of investigation (see, for instance, Montague and Berns, 2002). It has been suggested that activity of the dopamine system that is not associated with prediction errors for reward might be associated with ensuring appropriate exploration (Suri and Schultz, 1999; Suri, 2002; Kakade and Dayan, 2001, 2002). It is also possible that slower, but more tonic, activation of dopamine neurons is responsible for rather different effects (e.g., Howland et al., 2002), perhaps those associated with biasing the animal toward specific places or contexts or toward certain strategies in instrumental conditioning. Different populations of dopamine receptors in the striatum may contribute to quite different functions, particularly given the suggestion that D1/D4 receptors regulate activity in the direct striato-nigra pathway, whereas D2/D3 receptors regulate activity in the indirect pathway, projecting to the substantia nigra via the pallidum and subthalamic nucleus (Gerfen, 1992; Gerfen et al., 1990). It likely that these different populations of dopamine receptors contribute to quite distinct functions.

The circuitry that mediates the interactions between Pavlovian and instrumental conditioning is of considerable interest. Although these forms of learning may involve distinct systems and may even be subserved by quite distinct learning rules, clearly, Pavlovian cues can exert an excitatory effect upon instrumental performance. Generally, the associative strength of a CS predicts its impact on instrumental performance (Rescorla and Solomon, 1967; Rescorla, 1968) and, as such, Pavlovian-instrumental transfer appears to offer the most direct evidence (and, arguably, the best model) of a direct connection between reflexive and goal-directed learning systems. Currently, evidence suggests that the point of interaction is likely to involve connections between the shell and core of the nucleus accumbens. Thus, for example, in a recent, unpublished study, Corbit and Balleine found that asymmetrical AMPA lesions of shell and core abolish transfer, suggesting that the critical interaction is between AMPA-related processes in these structures. Van Dongen et al. (2001, Soc. Neurosci. Abstr.) has recently described direct and reciprocal projections from shell to core, and Haber et al. (2000) have described spiralling interactions between the striatum and midbrain dopamine neurons that support, for example, a shell-VTA-core interaction. In this case, shell projection neurons could exert quite direct control over plasticity in the accumbens core. It may be that a structure downstream from shell and core is the site of integration. The ventral pallidum is a reasonable possibility in this regard, particularly given that the GABA projection neurons from the striatum to the pallidum appear to be predominantly controlled by an AMPA-related process (Byrnes et al., 1997).

Various features of instrumental and Pavlovian conditioning have yet to be fully integrated into the current approach. With respect to Pavlovian processes, we have concentrated almost exclusively on excitatory appetitive conditioning. Inhibitory and aversive conditioning (and even extinction of Palovian excitors) pose a further set of important concerns. Unfortunately, there is just as little agreement about the neural underpinnings of aversive conditioning as there is of appetitive condition-

ing. One strongly favored possibility is that there is affective opponency (Solomon and Corbit, 1974; Grossberg, 1988), the dynamics of which offer quite powerful accounts of various experimental phenomena. Based on pharmacological, behavioral, anatomical, and physiological data, it has been suggested (e.g., Deakin, 1983; Deakin and Graeff, 1991; Daw et al., 2002) that serotonin released by neurons whose cell bodies live in the dorsal raphe nucleus acts as an aversive opponent to dopamine. Different timescales may be involved, with the phasic component of the serotonin signal reporting on aversive events, but with its tonic component reporting on long-run average reward. At present, there is limited direct evidence for such a proposition, although there is good evidence from behavioral studies to support the existence of an aversive motivational system standing in a mutually inhibitory relationship with the appetitive system of Figure 2 (cf. Dickinson and Balleine, 2002). For example, conditioned inhibitors of aversive USs can block appetitive conditioning to CSs that would otherwise predict reward (e.g., Dickinson and Dearing, 1979), suggesting that there must be some connection between appetitive and aversive motivational centers.

With respect to instrumental conditioning, the suggestion that dopamine plays a limited role in changes in the value of the instrumental outcome accomplished by incentive learning raises the question not only of what neural systems, but also what neurotransmitter systems, might generally support this learning and hedonic evaluation. In this regard, the endogenous opiate system stands out as a promising candidate. There is abundant evidence that animals find exogenous opiates rewarding and, although some evidence has implicated VTA dopamine activity in these effects (Di Chiara and North, 1992; Phillips et al., 1983), the direct involvement of dopamine in mediating the reinforcing effects of opiates is questionable. Thus, for example, morphine and heroin self administration persists even after much of the dopaminergic activity in the ventral striatum is abolished by 6-OHDA lesions (Dworkin et al. 1988; Pettit et al., 1984). There are a number of reasons for supposing that the incentive value of foods and fluids is mediated by activation of the endogenous opiate system. First, morphine administration increases food intake (Gosnell and Majchrzak, 1993; Pecina and Berridge, 2000) and both increases the incidence of ingestive taste reactivity patterns to sweet tastes and reduces the rejection reactions produced by bitter tastes (Clarke and Parker, 1995; Doyle et al., 1993). Similarly, the opiate antagonists naloxone and naltrexone reduce the ingestive reactivity responses (Parker et al., 1992; Kelley et al., 1996) and, in humans, have been reported to reduce ratings of the palatability of food (Drewnowski et al., 1992; Bertino et al., 1991) and to produce dysphoria (Martin del Campo et al., 1994). At a neural level, current evidence suggests that opiate receptors in the ventral pallidum are likely to prove critical to the way that the endogenous opiate system generally affects palatability reactions to foods and fluids and hedonic tone (Johnson et al., 1993; Gong et al., 1996; Pecina and Berridge, 2000) and, as such, future experiments directly assessing the involvement of the opiate system generally and of the pallidum in particular in incentive learning effects would be very informative.

A further set of experiments that would be most helpful for specifying the model more fully concern the truly instrumental status of conditioned reinforcement and its role in sequential decision making. Determining whether or not (Pavlovian) conditioned reinforcers can support truly instrumental behaviors is long overdue. If they can, then this implies a somewhat more direct relationship between Pavlovian and instrumental motivational control than strict readings of recent theories would suggest and poses an interesting problem about the neural realization of these two phenomena. The role of accumbens dopamine in this also merits further exploration.

Sequential decision problems form an important computational class, and the substantial theories of dynamic programming (also employed in studies of optimal foraging, Mangel and Clark, 1988) are devoted to their solution. It is not practical even for computers to perform extensive 'cognitive' action choice (called forward-chaining) in the contemplation of appropriate actions at a particular state in a complex sequential decision problem. Reinforcement learning comprises one family of methods for performing dynamic programming, using the values of states as a form of cache for the reward consequences of the actions subsequent to those states. The activity of dopamine cells in simple versions of such tasks suggests that monkeys indeed acquire such values but does not indicate how they are represented and how they might be integrated with the sort of more limited forward models that are implied by some of the experiments on incentive learning. It would be interesting (and only a small extension to existing work) to pose stochastic sequential decision problems (called Markov decision problems) to monkeys while recording dopamine activity and manipulating the motivational relevance of outcomes, as in incentive learning.

The close relationship between dopamine activity and temporal difference error in reinforcement learning laid bare a new and rich connection between animal and artificial decision making. The powerful logic of this seems to have acted as a blinder to the results of the sophisticated motivational experiments that we discussed here, that provide significant constraints on this connection. As in our admittedly preliminary proposal, taking note of these data should lead to richer and more psychologically and neurally faithful models.

Acknowledgments

We are very grateful to Christian Balkenius, Tony Dickinson, Sham Kakade, John O'Doherty, Emmet Spier, and Angela Yu for discussions. P.D. was funded by the Gatsby Charitable Foundation and B.W.B. by NIMH grant MH56446.

References

Alexander, G.E., and Crutcher, M.D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci. *13*, 266–271.

Baird, L.C. (1993). Advantage Updating. Technical report WL-TR-93-1146 (Dayton, OH: Wright-Patterson Air Force Base).

Balleine, B.W. (1992). The role of incentive learning in instrumental performance following shifts in primary motivation. J. Exp. Psychol. Anim. Behav. Process. *18*, 236–250.

Balleine, B.W. (1994). Asymmetrical interactions between thirst and

hunger in Pavlovian-instrumental transfer. Q. J. Exp. Psychol. B *47*, 211–231.

Balleine, B.W. (2001). Incentive processes in instrumental conditioning. In Handbook of Contemporary Learning Theories, R. Mowrer and S. Klein, eds. (Hillsdale, New Jersey: Erlbaum) pp. 307–366.

Balleine, B.W., and Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. Q. J. Exp. Psychol. B *43*, 279–296.

Balleine, B.W., and Dickinson, A. (1992). Signalling and incentive processes in instrumental reinforcer devaluation. Q. J. Exp. Psychol. B *45*, 285–301.

Balleine, B.W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology *37*, 407–419.

Balleine, B.W., and Dickinson, A. (2000). Effect of lesions of the insular cortex on instrumental conditioning: evidence for a role in incentive memory. J. Neurosci. *20*, 8954–8964.

Balleine, B.W., and Killcross, A.S. (1994). Effects of ibotenic acid lesions of the nucleus accumbens on instrumental action. Behav. Brain Res. *65*, 181–193.

Balleine, B.W., Ball, J., and Dickinson, A. (1994). Benzodiazepine-induced outcome revaluation and the motivational control of instrumental action. Behav. Neurosci. *108*, 573–589.

Barto, A.G., Sutton, R.S., and Anderson, C.W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics *5*, 834–846.

Berridge, K.C. (2001). Reward learning: reinforcement, incentives, and expectations. InThe Psychology of Learning and Motivation: Advances in Research and Theory, Volume 40, D.L. Medin, ed. (San Diego: Academic Press) pp. 223–278.

Berridge, K.C., and Robinson, T.E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res. Brain Res. Rev. *28*, 309–369.

Berridge, K.C., and Schulkin, J. (1989). Palatability shift of a salt-associated incentive during sodium depletion. Q. J. Exp. Psychol. B *41*, 121–138.

Bertino, M., Beauchamp, G.K., and Engelman, K. (1991). Naltrexone, an opioid blocker, alters taste perception and nutrient intake in humans. Am. J. Physiol. *261*, R59–R63.

Bindra, D. (1974). A motivational view of learning, performance, and behavior modification. Psychol. Rev. *81*, 199–213.

Bindra, D. (1978). How adaptive behavior is produced: a perceptual motivational alternative to response-reinforcement. Behav. Brain Sci. *1*, 41–52.

Bolles, R.C. (1975). Theory of Motivation (New York: Harper & Row).

Braun, J.J. (1990). Gustatory cortex: definition and function. In The Cerebral Cortex of the Rat, B. Kolb and R.C. Tees, eds. (Cambridge, MA: MIT Press), pp. 407–430.

Byrnes, E.M., Reilly, A., and Bruno, J.P. (1997). Effects of AMPA and D1 receptor activation on striatal and nigral GABA efflux. Synapse *26*, 254–268.

Cador, M., Robbins, T.W., and Everitt, B.J. (1989). Involvement of the amygdala in stimulus-reward associations: interaction with the ventral striatum. Neuroscience *30*, 77–86.

Clarke, S.N.D.A., and Parker, L.A. (1995). Morphine-induced modification of quinine palatability: effects of multiple morphine-quinine trials. Pharmacol. Biochem. Behav. *51*, 505–508.

Colwill, R.C., and Rescorla, R.A. (1986). Associative structures in instrumental learning. In The Psychology of Learning and Motivation, Volume 20, G.H. Bower, ed. (New York: Academic Press), pp. 55–104).

Colwill, R.M., and Motzkin, D.K. (1994). Encoding of the unconditioned stimulus in Pavlovian conditioning. Anim. Learn. Behav. *22*, 384–394.

Colwill, R.M., and Rescorla, R.A. (1988). Associations between the discriminative stimulus and the reinforcer in instrumental learning. J. Exp. Psychol. Anim. Behav. Process. *14*, 155–164.

Corbit, L., Muir, J., and Balleine, B.W. (2001). The role of the nucleus accumbens in instrumental conditioning: evidence for a functional dissociation between accumbens core and shell. J. Neurosci. *21*, 3251–3260.

Davidson, T.L., Altizer, A.M., Benoit, S.C., Walls, E.K., and Powley, T.L. (1997). Encoding and selective activation of "metabolic memories" in the rat. Behav. Neurosci. *111*, 1014–1030.

Davis, J., and Bitterman, M.E. (1971). Differential reinforcement of other behavior (DRO): a yoked-control comparision. J. Exp. Anal. Behav. *15*, 237–241.

Daw, N.D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. Neural Networks *15*, 603–616.

Dayan, P. (1993). Improving generalization for temporal difference learning: the successor representation. Neural Comput. *5*, 613–624.

Dayan, P. (2002). Motivated reinforcement learning. In Advances in Neural Information Processing Systems, 2001, T.G. Dietterich, S. Becker, and Z. Ghahramani, eds. (Cambridge, MA: MIT Press), in press.

Deakin, J.F.W. (1983). Roles of brain serotonergic neurons in escape, avoidance and other behaviors. J. Psychopharmacol. *43*, 563–577.

Deakin, J.F.W., and Graeff, F.G. (1991). 5-HT and mechanisms of defence. J. Psychopharmacol. *5*, 305–316.

DeBold, R.C., Miller, N.E., and Jensen, D.D. (1965). Effect of strength of drive determined by a new technique for appetitive classical conditioning of rats. J. Comp. Physiol. Psychol. *59*, 102–108.

de Borchgrave, R., Rawlins, J.N.P., Dickinson, A., and Balleine, B.W. (2002). The role of the nucleus accumbens in instrumental conditioning. Exp. Brain Res. *144*, 50–68.

Di Chiara, G., and North, R.A. (1992). Neurobiology of opiate abuse. Trends Pharmacol. Sci. *13*, 185–193.

Dickinson, A. (1994). Instrumental conditioning. In Animal Learning and Cognition, N.J. Mackintosh, ed. (San Diego: Academic Press), pp. 45–79.

Dickinson, A. (1997). Bolles's psychological syllogism. In Learning, Motivation, and Cognition: The Functional Behaviorism of Robert C. Bolles, M.E. Bouton, and M.S. Fanselow, eds. (Washington, DC: American Psychological Association), pp. 345–367.

Dickinson, A., and Balleine, B.W. (1993). Actions and responses: the dual psychology of behaviour. In Spatial Representation, N. Eilan, R. McCarthy, and M.W. Brewer, eds. (Oxford: Basil Blackwell Ltd.), pp. 277–293.

Dickinson, A., and Balleine, B.W. (1994). Motivational control of goal-directed action. Anim. Learn. Behav. *22*, 1–18.

Dickinson, A., and Balleine, B.W. (1995). Motivational control of instrumental action. Current Directions in Psychological Science *4*, 162–167.

Dickinson, A., and Balleine, B.W. (2002). The role of learning in motivation. In Learning, Motivation and Emotion, Volume 3 of Steven's Handbook of Experimental Psychology, Third Edition, C.R. Gallistel, ed. (New York: John Wiley & Sons), in press.

Dickinson, A., and Dawson, G.R. (1987). Pavlovian processes in the motivational control of instrumental performance. Q. J. Exp. Psychol. B *39*, 201–213.

Dickinson, A., and Dawson, G.R. (1988). Motivational control of instrumental performance: the role of prior experience of the reinforcer. Q. J. Exp. Psychol. B *40*, 113–134.

Dickinson, A., and Dearing, M.F. (1979). Appetitive-aversive interactions and inhibitory processes. In Mechanism of Learning and Motivation, A. Dickinson and R.A. Boakes, eds. (Hillsdale, NJ: Lawrence Erlbaum Associates), pp. 203–231.

Dickinson, A., and Mulatero, C.W. (1989). Reinforcer specificity of the suppression of instrumental performance on a non-contingent schedule. Behavioral Processes *19*, 167–180.

Dickinson, A., Balleine, B.W., Watt, A., Gonzalez, F., and Boakes, R.A. (1995). Motivational control after extended instrumental training. Anim. Learn. Behav. *23*, 197–206.

Dickinson, A., Squire, S., Varga, Z., and Smith, J.W. (1998). Omission learning after instrumental pretraining. Q. J. Exp. Psychol. B *51*, 271–286.

Dickinson, A., Smith, J., and Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. Behav. Neurosci. *114*, 468–483.

Doyle, T.G., Berridge, K.C., and Gosnell, B.A. (1993). Morphine enhances hedonic taste palatability in rats. Pharmacol. Biochem. Behav. *46*, 745–749.

Drewnowski, A., Krahn, D.D., Demitrack, M.A., Nairn, K., and Gosnell, B.A. (1992). Taste responses and preferences for sweet high-fat foods: evidence for opioid involvement. Physiol. Behav. *51*, 371–379.

Dworkin, S.I., Guerin, G.F., Goerders, N.E., and Smith, J.E. (1988). Lack of an effect of 6-hydroxydopamine lesions of the nucleus accumbens on intravenous morphine self-administration. Pharmacol. Biochem. Behav. *30*, 1051–1057.

Everitt, B.J., and Stacey, P. (1987). Studies of instrumental behavior with sexual reinforcement in male rats (Rattus novegicus): II. effects of preoptic area lesions, castration, and testosterone. J. Comp. Psychol. *101*, 407–419.

Fudim, O.K. (1978). Sensory preconditioning of flavors with formalin-produced sodium need. J. Exp. Psychol. Anim. Behav. Process. *4*, 276–285.

Gerfen, C.R. (1985). The neostriatal mosaic. I. compartmental organisation of projections from the striatum to the substantia nigra of the rat. J. Comp. Neurol. *236*, 454–476.

Gerfen, C.R. (1992). The neostriatal mosaic: multiple levels of compartmental organization. Trends Neurosci. *15*, 133–139.

Gerfen, C.R., Engber, T.M., Mahan, L.C., Susel, Z., Chase, T.N., Monsma, F.J., and Sibley, D.R. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. Science *250*, 1429–1432.

Gluck, M.A., Allen, M.T., Myers, C.E., and Thompson, R.F. (2001). Cerebellar substrates for error correction in motor conditioning. Neurobiol. Learn. Mem. *76*, 314–341.

Gong, W., Neil, D., and Justice, J.B. (1996). Conditioned place preference and locomotor activation produced by injection of psychostimulants into ventral pallidum. Brain Res. *707*, 64–74.

Gosnell, B.A., and Majchrzak, M.J. (1993). Centrally administered opioid peptides stimulate saccharin intake in nondeprived rats. Pharmacol. Biochem. Behav. *33*, 805–810.

Graybiel, A.M., Aosaki, T., Flaherty, A.W., and Kimura, M. (1994). The basal ganglia and adaptive motor control. Science *265*, 1826–1831.

Grossberg, S.E. (1988). Neural Networks and Natural Intelligence. (Cambridge, MA: MIT Press).

Haber, S.N., Fudge, J.L., and McFarland, N.R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J. Neurosci. *20*, 2369–2382.

Hammond, L.J. (1980). The effects of contingencies upon appetitive conditioning of free-operant behavior. J. Exp. Anal. Behav. *34*, 297–304.

Hendersen, R.W., and Graham, J. (1979). Avoidance of heat by rats: effects of thermal context on the rapidity of extinction. Learn. Motiv. *10*, 351–363.

Hershberger, W.A. (1986). An approach through the looking glass. Anim. Learn. Behav. *14*, 443–451.

Holland, P.C. (1977). Conditioned stimulus as a determinant of the form of the Pavlovian conditioned response. J. Exp. Psychol. Anim. Behav. Process. *3*, 77–104.

Holland, P.C. (1979). Differential effects of omission contingencies on various components of Pavlovian appetitive responding in rats. J. Exp. Psychol. Anim. Behav. Process. *5*, 178–193.

Holland, P.C. (1993). Cognitive aspects of classical conditioning. Curr. Opin. Neurobiol. *3*, 230–236.

Hollis, K.L., Pharr, V.L., Dumas, M.J., Britton, G.B., and Field, J. (1997). Classical conditioning provides paternity advantage for territorial male blue gouramis (Trichogaster trichopterus). J. Comp. Psychol. *111*, 219–225.

Houk, J.C., Adams, J.L., and Barto, A.G. (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, J.C. Houk and J.L. Davis, eds. (Cambridge, MA: MIT Press), pp. 249–270.

Howland, J.G., Taepavarapruk, P., and Phillips, A.G. (2002). Glutamate receptor-dependent modulation of dopamine efflux in the nucleus accumbens by basolateral, but not central, nucleus of the amygdala in rats. J. Neurosci. *22*, 1137–1145.

Johnson, P.I., Stellar, J.R., and Paul, A.D. (1993). Regional reward differences within the ventral pallidum are revealed by microinjections of a mu opiate receptor agonist. Neuropharmacology *32*, 1305–1314.

Kakade, S., and Dayan, P. (2001). Dopamine bonuses. In Advances in Neural Information Processing Systems 2000, T.K. Leen, T.G. Dietterich, and V. Tresp, eds. (Cambridge, MA: MIT Press), pp. 131–137.

Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. Neural Networks *15*, 549–559.

Kearns, M.S., and Singh, S. (1999). Finite-sample rates of convergence for Q-learning and indirect methods. In Advances in Neural Information Processing Systems, Volume 11, M. Kearns, S.A. Solla, and D. Cohn, eds. (Cambridge, MA: MIT Press), pp. 996–1002.

Kelley, A.E., Bless, E.P., and Swanson, C.J. (1996). Investigation of the effects of opiate antagonists infused into the nucleus accumbens on feeding and sucrose drinking in rats. J. Pharmacol. Exp. Ther. *278*, 1499–1507.

Konorski, J. (1967). Integrative Activity of the Brain: An Interdisciplinary Approach (Chicago: University of Chicago Press).

Krebs, J.R., Kacelnik, A., and Taylor, P. (1978). Test of optimal sampling by foraging great tits. Nature *275*, 27–31.

Krieckhaus, E.E. (1970). "Innate recognition" aids rats in sodium regulation. J. Comp. Physiol. Psychol. *73*, 117–122.

Lopez, M., and Paredes-Olay, C. (1999). Sensitivity of instrumental responses to an upshift in water deprivation. Anim. Learn. Behav. *27*, 280–287.

Lopez, M., Balleine, B., and Dickinson, A. (1992). Incentive learning and the motivational control of instrumental performance by thirst. Anim. Learn. Behav. *20*, 322–328.

Mackintosh, N.J. (1974). The Psychology of Animal Learning (New York: Academic Press).

Mangel, M., and Clark, C.W. (1988). Dynamic Modeling in Behavioral Ecology (Princeton: Princeton University Press).

Martin del Campo, A.F., Dowson, J.H., Herbert, J., and Paykel, E.S. (1994). Effects of naloxone on diurnal rhythms in mood and endocrine function: a dose response study. Psychopharmacology (Berl.) *114*, 583–590.

Mitchell, D.S., and Gormezano, I. (1970). Effects of water deprivation on classical appetitive conditioning of the rabbit's jaw movement response. Learn. Motiv. *1*, 199–206.

Montague, P.R., and Berns, G.S. (2002). Neural economics and biological substrates of valuation. Neuron *36*, this issue, 265–284.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. *16*, 1936–1947.

O'Brien, C.P., Childress, A.R., Ehrman, R., and Robbins, S.J. (1998). Conditioning factors in drug abuse: can they explain compulsion? J. Psychopharmacol. *12*, 15–22.

Parker, L.A., Maier, S., Rennie, S.M., and Crebolder, J. (1992). Morphine- and naltrexone-induced modification of palatability: analysis by the taste reactivity test. Behav. Neurosci. *106*, 999–1010.

Pecina, S., and Berridge, K.C. (2000). Opioid site in nucleus accumbens shell mediates eating and hedonic 'liking' for food: map based on microinjection Fos plumes. Brain Res. *863*, 71–86.

Pecina, S., Berridge, K.C., and Parker, L.A. (1997). Pimozide does not shift palatability: separation of anhedonia from sensorimotor effects. Pharmacol. Biochem. Behav. *58*, 801–811.

Pettit, H.O., Ettenberg, A., Bloom, F.E., and Koob, G.F. (1984). Destruction of dopamine in the nucleus accumbens selectively attenuates cocaine but not heroin self-administration in rats. Psychopharmacology (Berl.) *84*, 167–173.

Phillips, A.G., LePiane, F.G., and Fibiger, H.C. (1983). Dopaminergic mediation of reward produced by direct injection of enkephalin into the ventral tegmental area in the rat. Life Sci. *33*, 2505–2511.

Rescorla, R.A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. J. Comp. Physiol. Psychol. *66*, 1–5.

Rescorla, R.A. (1982). Simultaneous second-order conditioning produces S-S learning in conditioned suppression. J. Exp. Psychol. Anim. Behav. Process. *8*, 23–32.

Rescorla, R.A., and Solomon, R.L. (1967). Two-process learning theory: relationships between Pavlovian conditioning and instrumental learning. Psychol. Rev. *74*, 151–182.

Reynolds, N.J., Hyland, B.I., and Wickens, J.R. (2001). A cellular mechanism of reward-related learning. Nature *413*, 67–70.

Rolls, E.T. (1989). Information processing in the taste system of primates. J. Exp. Biol. *146*, 141–164.

Rolls, E.T. (2000a). The orbitofrontal cortex and reward. Cereb. Cortex *10*, 284–294.

Rolls, E.T. (2000b). Memory systems in the brain. Annu. Rev. Psychol. *51*, 599–630.

Rolls, E.T., Yaxley, S., and Sienkiewicz, Z.J. (1990). Gustatory responses of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. J. Neurophysiol. *64*, 1055–1066.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J. Neurophysiol. *80*, 1–27.

Schultz, W. (2002). Getting formal with dopamine and reward. Neuron *36*, this issue, 241–263.

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

Solomon, R.L., and Corbit, J.D. (1974). An opponent-process theory of motivation. I. Temporal dynamics of affect. Psychol. Rev. *81*, 119–145.

Suri, R.E. (2002). TD models of reward predictive responses in dopamine neurons. Neural Networks *15*, 523–533.

Suri, R.E., and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp. Brain Res. *121*, 350–354.

Suri, R.E., and Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience *91*, 871–890.

Suri, R.E., and Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. Neural Comput. *13*, 841–862.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (Cambridge, MA: MIT Press).

Sutton, R.S., and Pinette, B. (1985). The learning of world models by connectionist networks. In Proceedings of the Seventh Annual Conference of the Cognitive Science Society (Irvine, CA: Lawrence Erlbaum), pp. 54–64.

Taylor, J.R., and Robbins, T.W. (1984). Enhanced behavioral control by conditioned reinforcers following micro-injections of d-amphetamine into the nucleus accumbens. Psychopharmacology (Berl.) *84*, 405–412.

Toates, F.M. (1986). Motivational Systems (Cambridge, UK: Cambridge University Press).

Toates, F.M. (1994). Comparing motivational systems: an incentive motivation perspective. In Appetite: Neural and Behavioural Bases, C.R. Legg and D.A. Booth, eds. (New York: Oxford University Press), pp 305–327.

Wagner, A.R., and Brandon, S.E. (1989). Evolution of a structured connectionist model of Pavlovian conditioning (AESOP). In Contemporary learning theories, S.B. Klein and R.R. Mowrer, eds. (Hillsdale, NJ: Lawrence Erlbaum Associates), pp. 149–189.

Watkins, C.J.C.H. (1989). Learning from delayed rewards. PhD thesis, University of Cambridge, Cambridge, United Kingdom.

Wyvell, C.L., and Berridge, K.C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: enhancement of reward "wanting" without enhanced liking or response reinforcement. J. Neurosci. *20*, 8122–8130.

Wyvell, C.L., and Berridge, K.C. (2001). Incentive sensitization by previous amphetamine exposure: increased cue-triggered "wanting" for sucrose reward. J. Neurosci. *21*, 7831–7840.

Wolpert, D.M., and Ghahramani, Z. (2000). Computational principles of movement neuroscience. Nat. Neurosci. *3*, 1212–1217.

Yaxley, S., Rolls, E.T., and Sienkiewicz, Z.J. (1988). The responsiveness of neurons in the insular gustatory cortex of the macaque monkey is independent of hunger. Physiol. Behav. *42*, 223–229.