CHAPTER

21

Multiple Systems for Value Learning

Nathaniel D. Daw and John P. O'Doherty

OUTLINE

Introduction 393

Multiple Systems for Learning and Controlling Behavior 394
- *Reflexes and Pavlovian Learning* 394
- *Instrumental Conditioning: Habits and Goal-Directed Actions* 395

Computational Foundations of Multiple Learning Systems 398
- *Model-Based and Model-Free Learning* 398
- *World Models Versus Goals and Habits* 399
- *Why Two Instrumental Systems?* 400
- *Computational Approaches to Pavlovian Learning* 400

Multiple Neural Systems for Value Learning 401
- *Pavlovian Learning* 401
- *Instrumental Behavior: Habit Learning* 402
- *Goal-Directed Learning* 403

What is the Nature of the Interactions Between the Different Systems? 405
- *Pavlovian and Instrumental Interactions* 405

Conclusions: Alternative Systems, More Systems 406

References 408

INTRODUCTION

According to expected utility theory, choice is unitary by definition. For instance, a single scale mapping the objects of choice to utility or value is implicit in (indeed, formally equivalent to; see Chapter 1) a set of preferences over these objects, so long as those preferences satisfy some regularities such as *transitivity*.

Of course, such an abstract analysis does not speak directly to the mechanisms and processes that actually produce choices. At a process level, the notion that human and animal decisions are governed not by a single unitary controller, but rather by multiple, competing sub-systems, is pervasive throughout the history of psychology (Damasio, 1994; Dickinson, 1985; Freud, 1961; James, 1950). Similar frameworks have also become prevalent in neuroscience and behavioral economics (Balleine and Dickinson 1998; Balleine *et al.*, 2008; Daw *et al.*, 2005; Kahneman, 2003; Laibson, 1997; Loewenstein and O'Donoghue, 2004; Thaler and An, 1981; Weber and Johnson, 2009).

Although such multiplicity of control sits oddly with some theoretical perspectives, as we stress below, the brain is modular, and it evolved over time. Behavioral control mechanisms exist even in primitive organisms and are preserved, and augmented, in humans and other mammals. Also, such a multiplicity of choice mechanisms can be normatively justified even in more theoretical analyses, once computational considerations are taken into account. Although theory prescribes that a decision variable such as expected utility should take some particular value, exactly computing this value to guide choice is often laborious or intractable. In this case, as we will see, approximations may be preferable overall, and different approximations are more efficient in different circumstances.

In this chapter, we focus on a particularly crisply defined and well-supported version of the multiple systems framework, which has its roots in the behavioral psychology of animal learning (Balleine and Dickinson, 1998; Dickinson, 1985), and has more recently been extended to humans and to serve as the foundation for predominant neural and computational accounts of these functions (Balleine and O'Doherty, 2010; Balleine *et al.*, 2008; Daw *et al.*, 2005). The overarching theme of all this work is that a particular

© 2014 Elsevier Inc. All rights reserved.

behavior – such as a lever press by a rat – can arise in multiple different ways, which are dissociable psychologically, neurally, and computationally. In effect, these are different routes to a decision.

This framework details three learning systems that enable organisms to draw on previous experience to make predictions about the world and to select behaviors appropriate to those predictions. Since these different sorts of predictions all ultimately concern events relevant to biological fitness, such as rewards or punishments, they can also be thought of as different *forms of value*. The systems are: a *Pavlovian system* that learns to predict biologically significant events so as to trigger appropriate responses; a *habitual system* that learns to repeat previously successful actions; and a *goal-directed system* that evaluates actions on the basis of their specific anticipated consequences.

In this chapter, we will describe each of these learning processes, detailing their putative neuronanatomical and computational underpinnings. Also, we will describe situations under which these different systems might interact with each other, in a manner that can bias behavior in either adaptive or maladaptive ways. Furthermore, we will consider whether the three systems as traditionally outlined are sufficient to account for the full gamut of human behavior, or whether there might be additional systems. Finally, we will speculate on the relationship between the multiple *learning system* framework we outline here and other multiple systems theories not overtly related to learning.

MULTIPLE SYSTEMS FOR LEARNING AND CONTROLLING BEHAVIOR

Reflexes and Pavlovian Learning

In order to understand these different systems it is instructive to take a phylogenetic perspective. For all animals, it confers adaptive advantage to have mechanisms in place to alter behavior in response to environmental challenges and thereby increase the probability of survival. Perhaps the simplest such behaviors are reflexes. These are fixed, stereotyped behaviors automatically elicited by specific types of stimuli (Sherrington, 1906). Such stimuli do not require learning (over the lifetime of the organism) in order to come to elicit such responses, but rather have innate activating tendencies. Simple examples of such reflexes are the withdrawal reflex elicited after touching a hot surface, a startle reaction elicited by a loud bang, or the generation of a salivary response following the presence of food in the mouth. These reflexes are behaviors that have been shaped over the course of evolutionary history because they provide an adaptive solution to environmental challenges: it is useful to withdraw from hot surfaces so as to minimize tissue damage, it is advantageous to salivate in the presence of food so as to facilitate its consumption and digestion. Reflexive behaviors are simple to implement (e.g., by more or less directly coupling sensors to effectors with minimal computation in between), and accordingly they are found in even the simplest organisms such as in species of bacteria that show chemotaxis (Berg *et al.*, 1972) all the way up to humans.

Reflexes are by nature reactive, in that they are elicited only once a triggering stimulus is perceived. In many cases, however, it would more advantageous for an organism to be able to behave prospectively, in advance of a behaviorally significant event. For example, a flight reflex might help you to survive an encounter with a mountain lion, but will be more effective if you can flee in anticipation when the predator is likely to show up, as opposed to only reacting once it is right in front of you.

Pavlovian learning (see also Chapter 15) is a mechanism by which an animal can learn to make predictions about when biologically significant events are likely to occur, and in particular to learn which stimuli (e.g., in the case of mountain lions: roars or rustling of leaves) tend to precede them (Pavlov, 1927). Such predictions can then be coupled to the reflex mechanism, so that instead of responding exclusively in a reactive manner, the organism can elicit reflexive actions in anticipation of a biologically significant event. The standard laboratory model of such learning is based on Pavlov's (1927) findings that if a neutral stimulus (e.g., a bell) is repeatedly paired with the subsequent delivery of food, then that stimulus will also come to elicit salivation by virtue of its predictive relationship with the food delivery. The response depends on what sort of outcome is predicted: if a neutral stimulus is paired with the subsequent delivery of thermal pain, that stimulus will come to elicit a withdrawal reflex. Although some types of conditioned response are identical to the response elicited by the associated outcome (e.g., salivation to food), some other classes of Pavlovian conditioned reflexes, while still stereotyped, are distinct from those that occur in response to the predicted outcome (such as orienting to a visual cue predicting food as opposed to chewing in response to the food itself), perhaps reflecting the fact that the behavior that would be adaptive in preparation for the arrival of an event is sometimes different from that required following its onset (Konorski, 1948).

Pavlovian learning is known to be present in many invertebrates, including insects such as drosophila (Tully and Quinn, 1985), and even in the sea-slug (aplysia; Walters *et al.*, 1981), and also in vertebrates including humans (Davey, 1992).

As described, Pavlovian behaviors are more flexible than simple reflexes in that when to emit the behaviors is shaped by predictive learning, but they are also inflexible since the responses themselves are stereotyped. A related point is that the learned contingency that controls Pavlovian behavior is that between the stimulus and the outcome rather than that between the action and the outcome. That is, I salivate because I have learned something about the bell – that it predicts food – rather than something about salivation – e.g., that it makes food more palatable. Clearly, learning to take actions because they produce some desired outcome (to carry out some arbitrary action such as lever-pressing, because it produces food) is also highly advantageous, and more reminiscent of decision making as it is normally conceived in economic analyses. However, before turning to such learning, known as instrumental conditioning, we first discuss the evidence that Pavlovian behaviors are really produced by a stimulus-triggered reflex in the manner described. Although some Pavlovian responses (like salivation) have an obviously automatic character, others (such as approach or withdrawal) are more ambiguous: the form of the behavior itself thus does not unambiguously reveal the nature of the learning that produced it.

Accordingly, a raging debate in the animal learning field during the mid-twentieth century concerned the issue of whether Pavlovian learning processes really existed separate from instrumental processes, or whether all animal learning could be explained by one or the other sort of mechanism (Bindra, 1976; Mowrer, 1947). Clear evidence that putatively Pavlovian behavior really is driven by learning about stimulus–outcome relationships (versus instrumental learning about action–outcome relationships) comes from experiments in which these two sorts of contingencies are pitted against one another. For instance, it can be arranged that a bell predicts food, but the food is only delivered on trials when the animal does not salivate. In this way, the animal experiences the Pavlovian stimulus–outcome relationship, but never the (instrumental) action–outcome relationship. Nevertheless, animals do come to salivate reliably in this situation, and are unable to learn not to do so, despite the fact that this deprives them of food (Sheffield, 1965). This (and similar results for other behaviors like approach; Hershberger, 1986) supports the interpretation of these behaviors as a Pavlovian reflex.

Instrumental Conditioning: Habits and Goal-Directed Actions

If Pavlovian behaviors are not instrumental, it is also the case that instrumental behaviors are not Pavlovian. That is, animals can also learn to emit new behaviors that produce desired outcomes, and are sensitive to the action–outcome contingency. Consider lever-pressing for food. In principle, behavior that appears instrumental might arise due to a Pavlovian reflex. A rat might approach a stimulus (here, a lever) predictive of food, and thereby blunder into depressing the lever. But if the behavior changes in response to changes in the action–outcome contingencies that do not affect the stimulus–outcome relationship, then such behaviors cannot be explained as Pavlovian. For instance, animals can either learn selectively to press the same lever to the left, or to the right, depending which of those movements is programmed to produce food (Dickinson, 1996). Both such behaviors can not be explained away as inbuilt Pavlovian reflexes to the expectancy of food.

Early theories of instrumental conditioning described the learning process in terms of a simple mechanism which is again an elaboration of the stimulus-triggered reflex. Here, the idea is to learn new associations between stimuli and responses: effectively, wiring up new behaviors as reflexes (Hull, 1943; Thorndike, 1898). Such stimulus–response links were suggested to be shaped by a reinforcement rule that the early 20th century psychologist Edward Thorndike called the *Law of Effect*. He proposed that if a response was performed in the presence of a stimulus, and it led to "satisfaction" (e.g., reward), then its link would be strengthened, whereas stimulus–response links leading to "discomfort" would be weakened. Such stimulus–response learning mechanisms are likely widely present across vertebrate species. Some even argue that instrumental learning mechanisms are present in inverterbrates such as drosophila or aplysia (Brembs *et al.*, 2002; Cook and Carew, 1986), although the extent to which Pavlovian accounts for the putative instrumental behavior have been successfully ruled out in some of the inverterbrate studies might be open to debate.

Stimulus–response learning of this sort is today referred to as *habitual learning* (Dickinson, 1985), and (after Pavlovian learning) is the second of the three behavioral control systems considered in this chapter. Although such a learning system is capable of establishing even very complex behavioral patterns, such as for instance when training a pigeon to play ping pong (Skinner, 1962), this mechanism still has an odd and sometimes maladaptive inflexibility owing to its foundation in the stimulus–response reflex. In particular, a stimulus–response learner works simply by repeating actions that were previously successful (i.e., followed by "satisfaction"). But such a mechanism is incapable of evaluating novel actions (or re-evaluating previously experienced ones) based on any other

information about the task, the world, or the animal's goals.

Based on this insight, in classic work, Thorndike's contemporary, the American psychologist Edward Tolman (Tolman, 1948) used a number of different spatial foraging tasks in the rat to argue for the insufficiency of the stimulus–response mechanism for explaining mammalian instrumental learning. For instance, he demonstrated that rats exposed to a maze, even in the absence of any reinforcement, were faster at learning the route to a location subsequently baited with food, compared to animals that hadn't been pretrained in the maze. This effect is known as *latent learning*. Similarly, Tolman demonstrated that rats could flexibly select new pathways through a maze in order to reach a goal if the previously rewarded pathway was no longer available or if a better shortcut newly became available. None of these effects can be explained by stimulus–response learning. For example, even if animals in the latent learning task formed some stimulus–response associations during the maze pre-exposure period, these wouldn't preferentially favor the particular trajectory that would later lead to reward. Tolman interpreted these findings as suggesting that these animals instead learned to encode what he called a "cognitive map" – in this case, essentially an internal map of the spatial layout of the maze and the locations of goals – and that they could use it in order flexibly to select actions in pursuit of their goals. More generally, a cognitive map (as defined today) typically encodes the contingencies of a task: how different actions lead to different outcomes, including goals.

More recently, Dickinson (1985) has argued (on the basis of a task and results discussed in more detail below) that both Thorndike and Tolman were, in effect, right: that in fact there are two distinct mechanisms for instrumental conditioning in the mammalian brain. These include both the habitual stimulus–response mechanism and a *goal-directed* mechanism that evaluates actions more prospectively, as by a cognitive map. The goal-directed system is the third system of behavioral control considered in this chapter. On Dickinson's definition, a choice is goal-directed if it depends on a representation of the action–outcome contingency (that lever-pressing produces food: the cognitive map) and on the outcome as a desired goal or incentive (that food is valuable). Otherwise it is seen as the product of some other influences, such as habitual or Pavlovian. As we have seen, Pavlovian and habitual mechanisms can produce adaptive behaviors, but they do not actually do so on the basis of a representation of this sort, the critical feature of a goal-directed system. For this reason, a goal-directed system can solve action selection problems that a habitual, stimulus–response system cannot, and this also allows its contributions to behavior to be dissociated experimentally from the products of habitual and Pavlovian systems.

A key basis for distinguishing goal-directed and habitual behaviors experimentally is thus examining whether organisms can flexibly adjust their actions following a change in the reward value of an associated outcome (Figure 21.1). In a typical experiment, a hungry rat first learns to lever-press for food. Following this training, some "devaluation" manipulation is performed to reduce the desirability of the food to the rat. For instance, the rat can be fed to satiety, or the food can be paired with drug-induced illness to produce a selective aversion. At this point the rat does not value the food, in the sense that it will not eat it if presented. The rat is then offered the chance to work again on the lever associated with the now-devalued food. In this case, if its behavior were controlled by a goal-directed system, then it would evaluate the action in terms of its consequence (the food) and its desirability (low), and correctly decide not to press the lever. This makes sense, but it stands in contrast to how a stimulus–response learner would behave. In that case, the behavior would only be controlled by its previous "satisfaction" upon pressing the lever. Because the act of pressing the lever was previously reinforced, such a system makes the counterintuitive prediction that the rat would continue to lever-press, even though it demonstrably doesn't want the food. This is because a stimulus–response system bases its choices only on past satisfaction, and not on the particular expected consequences of actions or their current values. Ultimately, the stimulus–response link can be unlearned (by new experience showing that the lever-press no longer produces "satisfaction"), but initially the mechanism will produce inappropriate behavior.

Dickinson and colleagues have used this reward devaluation manipulation, changing the value of an outcome after learning, to examine what mechanism drives instrumental behavior in rats. These experiments demonstrate evidence for both goal-directed and habitual control, i.e. rats can be either sensitive to devaluation (reducing pressing on a lever than had delivered now-devalued food, relative to a control non-devalued action–outcome pair), or insensitive (maintaining inappropriate lever pressing), in different circumstances. Note that much as with Tolman's spatial experiments, correctly adjusting one's action preferences following a reward devaluation cannot be explained by a stimulus–response mechanism. Conversely, a true goal-directed system would not persist in lever pressing for a devalued goal. These two modes of behavior thus each reject one of the models,

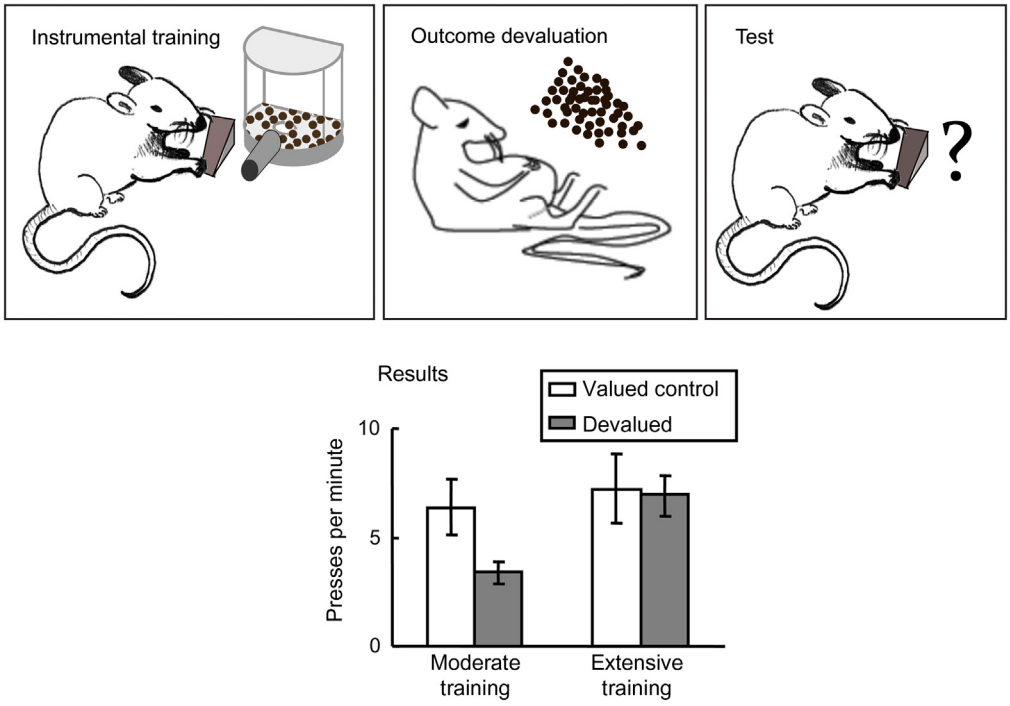

FIGURE 21.1 Distinguishing habitual from goal-directed instrumental learning using outcome devaluation. Left: Rats are first trained, when hungry, to lever press for food. Center: the food is devalued, e.g., by feeding the animal to satiety. Right: animals are tested to assess whether they will maintain lever pressing for the devalued outcome, compared to control animals who still value the outcome. The test is conducted in extinction (without food delivery) to ensure that any changes in behavior relate to the animal's internal representation of the outcome, rather than learning from new experience with it during the test phase. *Drawings by Sara Constantino*. Bottom: both devaluation-sensitive (goal-directed) and devaluation-insensitive (habitual) responses are observed under different circumstances. In this graph, the effect of the amount of instrumental training is illustrated (*data replotted from Holland, 2004*): goal-directed behavior dominates early, but gives rise to habitual (devaluation-insensitive) behavior following overtraining.

and suggest that two systems for instrumental conditioning control behavior at different times.

What circumstances affect which behavior is observed? One key factor among several that influence whether behavior is goal-directed or habitual is the amount of training the animal received in the initial lever-pressing, prior to devaluation (Adams, 1982; Balleine and Dickinson, 1998). If animals are moderately trained on the original lever-pressing task (e.g., having practiced it for five half-hour sessions prior to the devaluation), they maintain devaluation sensitivity; if they are overtrained (e.g., 20 sessions of lever-pressing practice), behavior can become insensitive to devaluation, suggesting a shift in control from goal-directed to habitual over the course of learning. This is one of the reasons for calling stimulus–response behavior *habitual* – it resonates with a phenomenon familiar in our own lives, that highly practiced behaviors (e.g., your route to work) become somehow automatic and can sometimes be performed even when inappropriate to your current goals (e.g., when you are actually headed to the grocery store). Recently, Tricomi and colleagues (2009) used a devaluation paradigm similar to the rodent studies to show a transition, with overtraining, from goal-directed to habitual behavior in humans as well.

Returning to our phylogenetic narrative, given that goal-directed and habitual behaviors are present in rodents as well as humans, it is certainly tempting to speculate that the capacity for goal-directed control might well be widespread among mammals. The phylogenetic viewpoint allows us to appreciate that as modern humans, we have inherited multiple different systems for interacting with and learning about the world, ranging from simple reflexes, including a Pavlovian controller, a habit system and then a goal-directed controller. On a behavioral level, these different control systems are likely to interact either cooperatively or competitively to guide behavior, sometimes in a manner that leads to the selection of apparently inappropriate behaviors. We next consider computational and neural approaches to this multiplicity.

COMPUTATIONAL FOUNDATIONS OF MULTIPLE LEARNING SYSTEMS

So far, we have presented behavioral evidence that several different sorts of learned representations can control behavior in humans and laboratory animals. We have suggested that these might reflect a series of increasingly sophisticated action control mechanisms built up by evolution, but we have so far been relatively informal in discussing how they are adaptive.

A complementary view on adaptive behavior comes from computer science and engineering, where researchers have considered the computational problem of learned optimal control in the context of controlling artificial agents such as robots. As discussed in Chapters 15 and 16, ideas and algorithms from this field, called reinforcement learning (RL; Sutton and Barto, 1998), are also influential as theories in computational neuroscience. In particular, because they focus on step-by-step optimization computations, such theories serve as a bridge between normative decision-theoretic or economic notions of adaptive behavior and the more process- or mechanism-level concerns of psychologists and neuroscientists.

As it turns out, the distinction between habitual and goal-directed instrumental control has a formal counterpart in RL, which we describe here (Figure 21.2A). This serves to connect the psychological categories to more abstract models of efficient behavior, and also to situate them in the context of studies of the neural mechanisms for this learning, which (as discussed in Chapters 15 and 16, and also below) have also been to a great extent understood in RL terms.

Model-Based and Model-Free Learning

Consider the problem of choosing among a set of options the one that maximizes the expected utility of the outcome. We could write a standard decision-theoretic expression for that objective, but using notation drawn from RL:

$$Q(a) = \sum_s P(s|a)r(s) \tag{21.1}$$

The options (actions a) result stochastically in different outcomes (outcome states s) which have different subjective utilities (rewards r); the objective function is the average reward in expectation over the outcome,

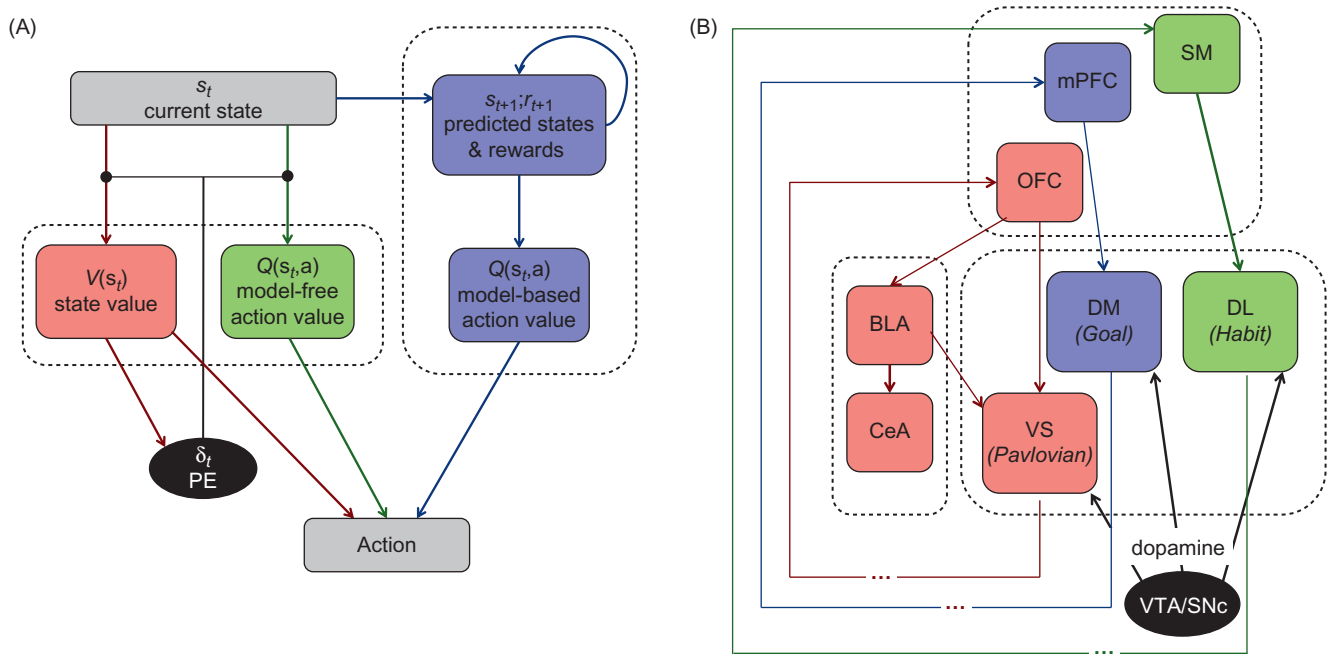

FIGURE 21.2 (A) Multiple routes to behavior in model-based and model-free reinforcement learning. (B) Neural circuits underlying valuation in conditioning. Areas are labeled based on rodent lesion studies; the identification of the homologous structures in primates is discussed in the text and illustrated in Figures 21.3 and 21.4. Evidence reviewed in text suggests that actions are influenced by three value learning systems implemented in dissociable neural substrates, each involving loops through different parts of the basal ganglia. On this view, habitual instrumental actions are encoded in loops involving sensory-motor (SM) cortical inputs to dorsolateral striatum (DL). A parallel circuit linking medial prefrontal cortex (mPFC), dorsomedial striatum (DM) appears to support goal-directed instrumental behavior. Finally, Pavlovian responses appear to involve a ventral loop linking orbitofrontal cortex (OFC) and ventral striatum (VS), with important contributions also of the central (CeA) and basal/lateral nuclei of the amygdala (BLA). All three loops are innervated by dopaminergic inputs from ventral tegmental area/substantia nigra pars compacta (VTA/SNc).

by convention written Q. Such a formalism is often used to characterize the choice between different monetary lotteries in a human decision task, but it is equally applicable, for instance, to a task in which a rat faces a number of different levers, which deliver different outcomes (e.g., different foods and liquids, shocks) according to different probabilities.

Now suppose, like the rat, we wish to *learn* to solve this problem by trial and error: by trying different actions and observing their results. A key insight in the early development of RL was that this learning problem could equally be attacked by focusing on estimating the quantities appearing on either side of the equal sign in Equation 21.1.

An approach based on the right hand side of the equation would learn a representation, for each action, of the likelihood of producing each outcome (i.e., $P(s|a)$), and also a representation of how pleasurable each outcome is (i.e., $r(s)$). These functions can be estimated from experience with actions and their outcomes, e.g., by counting and averaging. Then whenever one wants to choose an action, each candidate's value can be explicitly computed via plugging the estimates into Equation 21.1 and taking the sum. This is called model-based reinforcement learning, because it centers on representing the two functions characterizing outcomes and their values, which are together known as an "internal model" of the task.

In fact, this is not the way of solving the problem that is most prominent in psychological or neuroscientific theories. An alternative is to learn to approximate the left-hand side of the equation directly (Sutton, 1988). Here, that amounts to maintaining a representation of the estimated expected value Q for each action. By simply sampling actions, and maintaining a running average of the rewards obtained, one can estimate Q directly, and eschew any intervening representation of the outcomes themselves. This approach is known as model-free RL: it does not rely on an internal model of the task contingencies. This is also exactly the learning approach detailed in Chapter 15: Q can be updated using error-driven learning (increased or decreased depending whether the reward is larger or smaller than expected). As also discussed there, this is also the approach associated with prominent accounts of the responses of dopamine neurons, which appear to carry such a prediction error signal for reward (Barto, 1995; Montague *et al*., 1996; Schultz *et al*., 1997).

World Models Versus Goals and Habits

Note how closely the model-based and model-free approaches mirror, respectively, the goal-directed and habitual learning mechanisms described in the previous section (Daw *et al*., 2005). For model-based RL, the two pieces of the internal model ($P(s|a)$ and $r(s)$) mirror the defining antecedents of the goal-directed response, the action–outcome contingency and the outcome's incentive value. Meanwhile, like a stimulus–response habit, the model-free value $Q(a)$ measures the previous reward obtained for a without reference to the outcome identity: it is, in effect, a record of previous "satisfaction" following a. (Note that for simplicity, Equation 21.1 is written as a function over actions in a single situation. One could alternatively define Q as itself also dependent on the state in which the action is taken: $Q(s,a) = \sum_{s'} P(s'|s,\ a)r(s')$. This characterizes the value for actions taken in different states or situations, in which case it more directly resembles the strength of the stimulus–response association for stimulus s and response a, in terms of the outcome s'.)

Model-based and model-free RL make analogous predictions about reward devaluation tasks as do the goal-directed and habitual mechanisms they formalize (Daw *et al*., 2005). If an action is trained for some outcome, and then the outcome is devalued, model-based RL will incorporate that devaluation into its computation of $Q(a)$ via $r(s)$ in Equation 21.1 and adjust behavior. Conversely, because model-free RL does not compute $Q(a)$ in terms of the outcome identity, it cannot adjust the learned value following devaluation but must relearn $Q(a)$ from additional experience with the outcome's new value.

Finally, although Equation 21.1 describes tasks involving a single action for a single outcome, multistep decision tasks, in which a series of states, actions, and rewards occur in sequence, are important both in computer science (for example chess) and in psychology (mazes). As discussed in Chapters 15 and 16, the same two families of RL approaches generalize straightforwardly to multistep decision problems. In this case, the relevant notion of the expected value for an action in a state (a board position in chess or a location in a maze) sums accumulated rewards over the series of states that would subsequently be encountered. Thus, as detailed in Chapter 16, the expression analogous to the right-hand side of Equation 21.1 for such a task must take the expectation not just over the outcome state s, but over the whole series of states following it, summing rewards over the resulting potential trajectories. Model-free learning of values can then be accomplished by temporal-difference methods of the kind discussed in Chapters 15 and 17, which generalize error-driven sampling of obtained rewards to the long-run cumulative reward. As described in Chapter 15, this is the algorithm typically associated with the dopamine response. As detailed in Chapter 16, model-*based* RL in this setting depends on

learning the sequential transition function $P(s'|s,a)$ describing how an action in a state leads to the next state, and then iterating it repeatedly to predict trajectories in computing the cumulative reward analogous to the right hand side of Equation 21.1. This allows the same theory to characterize spatial tasks, cognitive maps and Tolman's results such as latent learning, along with simpler nonsequential operant lever-pressing tasks of the sort discussed above. Computing expected action values via model-based RL in this case resembles a sort of predictive simulation about what series of states and rewards will follow a choice (Johnson and Redish 2007; Schacter *et al.*, 2007).

Why Two Instrumental Systems?

The computational approach also provides some insight into two questions central to this chapter: why should the brain employ two *instrumental* decision systems, and how can it arbitrate between them? One answer is that these computational approaches represent different tradeoffs between computational costs and statistically efficient learning.

A model-based method is computationally expensive at decision time, since it must compute the expected value from Equation 21.1 explicitly, summing over different possible state sequences. This is a particularly onerous requirement in a sequential decision-making task like the game of chess, where the number of future board positions to be examined is typically impossibly large and the computation is therefore both laborious and approximate. Model-free RL, in contrast, requires only retrieving and comparing the learned net values Q at decision time. On the other hand, as we have seen, model-free RL can under certain circumstances make decisions that are less than ideal with respect to the agent's current knowledge and goals, as in the example of a rat working for devalued food. This is an example of a more general shortcoming of these algorithms, which is that the process for sampling Q values directly, without building a model, does not fully take into account all information available at any particular point that is relevant to estimate an action's value. By recomputing action values from their elements at each step, model-based RL ensures more efficient use of information.

These models thus clarify how each of these two approaches, model-based and model-free or goal-directed and habitual learning, have strengths and weaknesses that trade off against each other. Essentially, they offer two points on a tradeoff between computational complexity and statistical accuracy, with model-based RL computing a reliable result with difficulty, and model-free RL offering an easier shortcut to a potentially less accurate result. Consider the example of rodent lever-pressing, which starts out goal-directed and becomes habitual with overtraining. (Data do not yet exist, incidentally, to determine whether this transition is abrupt or gradual.) Early in training, the animal's experience is sparse, and it may be worthwhile to squeeze the most information out of this experience by laboriously computing the value of actions in a model-based fashion. However, given extensive experience lever-pressing in a stable environment, recomputing the same values in this way is unlikely to reveal any surprises. These sorts of considerations may thus help to explain why the brain apparently implements both methods, and when (or even how) it chooses one over the other. Formal models extend this reasoning (Daw *et al.*, 2005; Keramati *et al.*, 2011; Simon and Daw, 2011a) to analyze under what circumstances the computational costs of model-based RL (for instance, in the brain, the opportunity cost of time and the caloric cost of firing neurons) are likely to be justified in terms of producing better decisions, i.e., those that ultimately earn more rewards. The theories predict the effect of overtraining on habits together with a number of other factors that also have been shown to affect this balance in experiments.

Computational Approaches to Pavlovian Learning

We have focused on computational accounts of instrumental conditioning, in an attempt to capture the distinction between goal-directed and habitual learning. What of our other behavioral control system, the Pavlovian one?

As we have seen, the different forms of instrumental conditioning can be associated with different methods for predicting $Q(s, a)$, the reward expected for an action in some situation. Computed different ways, we can see this as playing both the role of the stimulus–response association (between s and a) or the goal-directed value of a in s. RL algorithms can analogously be applied to learn a similar function, which is relevant to Pavlovian conditioning in that it captures something like the stimulus–outcome association. This is known as $V(s)$, and represents the reward expected following some state (stimulus or situation), regardless of (in expectation over) any actions taken. As is described in detail in Chapter 15, model-free RL algorithms for predicting V using error-driven updating have a long history in psychology as theories of Pavlovian conditioning. Moreover, as also discussed below, these theories also became the foundation for accounts of the dopamine response and its involvement in conditioning.

For Pavlovian behavior, the assumption is that the conditioned response (salivation, or the like) in some state s is directly proportional to the predicted reward $V(s)$. Theoretical work has considered how these responses can compete against or combine with instrumental ones produced by model-free or model-based RL (Dayan *et al.*, 2006), but there has been little work attempting to understand or rationalize the principles of these interactions, analogous to that investigating the efficient tradeoff between model-based and model-free methods. This is an important area in which future work can be expected to yield significant results.

MULTIPLE NEURAL SYSTEMS FOR VALUE LEARNING

We have identified multiple learning systems that are dissociable behaviorally and operate according to different computational principles. This research has provided the foundation for seeking neural substrates supporting these mechanisms. This work is important for several reasons. Perhaps most crucially with respect to the argument of this chapter, the existence of dissociable neural substrates mediating these different behaviors supports their interpretation in terms of multiple systems. Second, information about the neural systems involved provides additional insights into how these systems operate. Third, investigations of the neural systems supporting these behaviors, many of which are interconnected, have tended to highlight additional questions about the nature of interactions and commonalities between the systems.

Broadly, our three learning systems implicate three adjacent subregions of the rodent striatum: ventral (for Pavlovian, also called the *ventral striatum* in primates including humans), dorsolateral (for habitual, called the putamen in primates) and dorsomedial (for goal-directed behaviors, called the caudate in primates; Figure 21.2B). These areas are interesting because they are all targets of the midbrain dopaminergic system (which plays an important role in computational accounts of RL), and because different areas of striatum have reciprocal interconnections with distinct areas of cortex via a series of "loops" through the basal ganglia (Alexander and Crutcher, 1990).

Pavlovian Learning

Pavlovian learning is arguably more heterogeneous than the other systems we have considered, since it involves numerous different sorts of responses for different predictions, many of which may involve distinct brain subsystems if only for expression. However, focusing on general appetitive and aversive Pavlovian conditioning procedures most relevant to neuroeconomics, there is now considerable evidence to implicate several brain structures in this process, particularly the amygdala, the ventral striatum and the orbitofrontal cortex. Anatomically, both the central nucleus of the amygdala and the ventral striatum are appropriately positioned to control the expression of different sorts of Pavlovian responses. The amygdala central nucleus projects to lateral hypothalamic and brainstem nuclei involved in implementing conditioned autonomic reflexes (Price and Amaral, 1981), while the ventral striatum sends projections via the globus pallidum to motor nuclei in the brain stem such as the peduncolopontine nucleus (Groenewegen and Russchen, 1984). These projections are compatible with a role for the ventral striatum in implementing conditioned skeletomotor reflexes such as approach and avoidance behavior, as well as consummatory responses such as licking. These areas and also areas upstream from them – the amygdala's basal and lateral nuclei and the orbitofrontal cortex – are all likely sites for plasticity subserving Pavlovian learning.

Accordingly, lesions of the whole amygdala and selective lesions of its lateral and central nuclei impair the acquisition and expression of aversive fear conditioning in rodents (Pare *et al.*, 2004). Lesions of the amygdala, ventral striatum and orbitofrontal cortex can all result in impairments in at least some forms of appetitive Pavlovian conditioning, such as conditioned approach (Hatfield *et al.*, 1996; Ostlund and Balleine, 2007; Parkinson *et al.*, 1999). Importantly, lesions of these areas tend not to have comparable effects on instrumental learning – supporting the dissociation of these functions – though (as we will discuss more below), lesions do implicate ventral striatum and basolateral amygdala in interactions between Pavlovian and instrumental mechanisms.

Single unit studies in both rodents and monkeys have revealed neuronal activity in both the amygdala and orbitofrontal cortex related to conditioned stimuli associated with the subsequent presentation of both appetitive and aversive unconditioned stimuli such as a sweet taste (juice reward), aversive taste or an air puff (Morrison and Salzman, 2011; Paton *et al.*, 2006; Schoenbaum *et al.*, 1998). Furthermore, human imaging studies have revealed neural responses in amygdala, ventral striatum and orbitofrontal cortex in response to conditioned stimuli that are predictive of the subsequent delivery of appetitive and aversive outcomes such as tastes and odors (Gottfried *et al.*, 2002, 2003; O'Doherty *et al.*, 2002).

An important commonality of the aforementioned areas is that they are all major targets of the dopamine-containing neurons of the midbrain, an observation that links these systems closely to the computational learning mechanisms discussed in the previous section. In particular (see also Chapter 15), the responses of dopamine neurons in Pavlovian conditioning experiments (and also dopamine release in ventral striatum assessed using fast-scan cyclic voltammetry) quantitatively match a reward prediction error signal that drives learning in model-free RL theories of Pavlovian conditioning (Montague *et al*., 1996; Schultz *et al*., 1997). Dopamine influences plasticity at its targets (notably, in striatum), which may ultimately subserve at least some forms of appetitive Pavlovian learning (Aggarwal *et al*., 2012; Reynolds and Wickens, 2002). Accordingly, dopamine in ventral striatum is indeed implicated in appetitive Pavlovian conditioning procedures (Parkinson *et al*., 2002). Importantly, however, dopamine (albeit in other areas of striatum) is also implicated in instrumental learning, as discussed below.

Instrumental Behavior: Habit Learning

The existence of two systems for instrumental behavior is supported by clear dissociations between brain networks across studies using a number of different methodologies. One approach combines lesions in rodents with devaluation tasks of the sort discussed above. The general form of the findings is that lesioned animals can acquire instrumental behaviors, but the reward devaluation test demonstrates that the behavior is supported by one or the other of goal-directed or habitual mechanisms (depending which areas are damaged) even under circumstances when the other system would dominate in control animals. Thus, for instance, following overtraining, animals with damage to areas involved in habits (discussed next) retain devaluation sensitivity even while the behavior of neurologically intact control animals become habitual (Yin *et al*., 2004). Another set of lesions preserves goal-directed behavior while abolishing habits.

These converging lines of evidence implicate the dorsolateral striatum (the rodent homologue of the putamen in primates) in habit learning and the habitual control of behavior. In rodents, lesions of the dorsolateral striatum have been found to render behavior permanently goal-directed such that even after overtraining these animals fail to express habits (Yin *et al*., 2004). These areas of striatum are connected to "skeletomotor loops" linking the basal ganglia with the motor cortices. Since these circuits are likely involved in the control of simple movements, these areas are well positioned for simple stimulus–response triggering (Alexander and Crutcher, 1990).

In humans, fMRI studies of motor sequence learning have reported an increase in activity in the posterior part of the dorsolateral striatum as sequences have become well learned (Jueptner *et al*., 1997; Lehéricy *et al*., 2005), although such studies typically have not determined whether responding has transitioned to becoming habitual using the appropriate behavioral assays. Tricomi *et al*. addressed those shortcomings by demonstrating that increasing activity in right posterolateral striatum over the course of training did relate to the emergence of habitual control as assessed with a reinforcer devaluation test (Tricomi *et al*., 2009) (Figure 21.3A,B). Moreover it has recently been shown that using diffusion tensor imaging (DTI: see Chapter 6) that differences in the strength of the connectivity between right posterolateral striatum and premotor cortex across individuals is significantly correlated with the degree to which individuals show evidence of habitual behavior on a task in which goal-directed and habitual responding are put in conflict with each other (de Wit *et al*., 2012). Finally, in a decision-making study (Wunderlich *et al*., 2012) based more on the computational distinction between model-based and model-free RL, correlates of value were seen in this region for extensively trained actions, but not for values that had to be computed by model-based search (Figure 21.3C).

From a computational perspective, instrumental learning of action preferences by model-free RL (the proposed computational theory of habit formation) uses reward prediction errors similar (or in some theories, identical) to those previously discussed for Pavlovian learning. Moreover, within instrumental conditioning, we might on computational grounds expect dopamine to be preferentially involved in habit formation, rather than in goal-directed behavior. This is because model-based learning (the putative computational substrate for goal-directed behavior) does not rely on similar reward prediction error signals, since it doesn't directly learn aggregate reward predictions at all (Glascher *et al*., 2010). Accordingly, attenuating prediction-error related signals in rodent dopamine neurons (by genetically deleting an excitatory receptor that supports such firing) impairs habits while sparing goal directed learning (Wang *et al*., 2011). Given the devaluation work discussed above, a likely site of action for dopaminergic involvement in habit learning is the dorsolateral striatum, where indeed the removal of the dopaminergic input blocks habit formation (Faure *et al*., 2005).

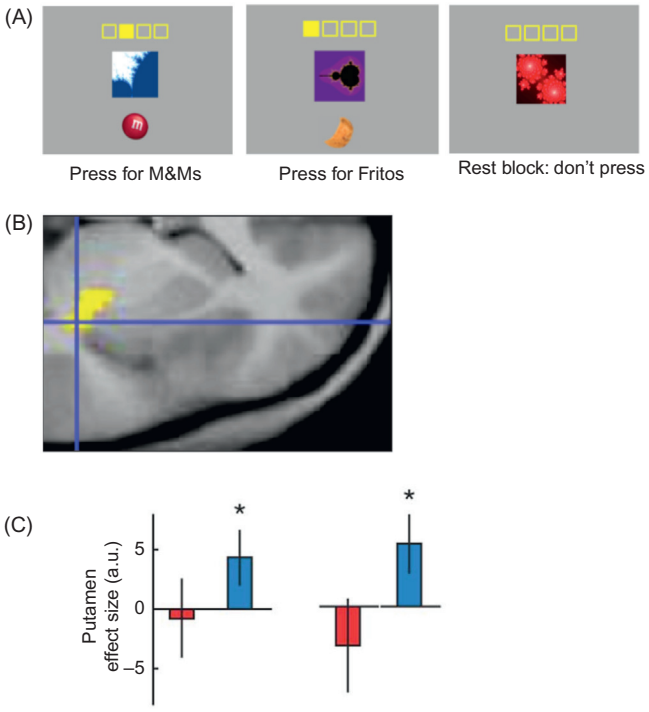

FIGURE 21.3 Neural correlates of putative habitual control and of model-free cached value signals in humans. (A) Task used by Tricomi and colleagues (2009) to induce habitual control in humans. Participants responded in an adapted free operant procedure whereby in a particular epoch they could free respond (on a variable interval schedule) in order to obtain rewarding outcomes which were either Fritos or M&Ms. The particular outcome obtained was conditional on performance of a particular action (button press) in the presence of a specific discriminative stimulus (fractal). One group of participants were extensively trained on these actions by being given 3 days of training (32 minutes per day). After devaluation by selective satiation on either Fritos or M&Ms, this group did not show any significant change in their response rates to the action associated with the devalued outcome relative to the non-devalued outcome in a test-phase, whereas by contrast another modestly trained group that received only 16 minutes of training on a single day showed robust devaluation effects. Thus, the extensively trained group showed evidence of habitual control. (B) Region of posterior putamen found to show gradual increases in activity as a function of training over 3 days of repeated scanning in the overtrained group, implicating this region in habitual control in humans. *From Tricomi* et al. *(2009).* (C) Results from fMRI study by Wunderlich and colleagues (2011) in which participants received extensive training on a particular sequence of actions in order to obtain rewards while at the same time they received modest training on another set of actions. In a choice situation between these two sets of actions, correlations were found in the identical region of putamen to that identified in C to the value of the extensively trained action but not to the moderately trained action irrespective of whether the modestly trained action is chosen (left plot), or whether the extensively trained action is chosen (right plot). *Adapted with permission from Wunderlich* et al. *(2012).*

Goal-Directed Learning

The counterpart to the lesion evidence implicating dorsolateral striatum in goal-directed behavior is lesion of the adjacent dorsomedial striatum in rodents, a manipulation which impairs goal-directed behavior as assessed by devaluation, but spares habits (Yin *et al.*, 2005). Together, these lesions of the dorsomedial and dorsolateral striatum suggest that these two functions are separate and independent from one another, in that they can each be separately affected while leaving the other intact. Such a pattern of results is known as a *double dissociation*.

Key inputs to the dorsomedial striatum come from the medial prefrontal cortex, and in particular, in rodents, from an area of the medial prefrontal cortex known as the prelimbic cortex. Lesions here have have effects similar to dorsomedial striatum in abolishing goal-directed behavior (Balleine and Dickinson, 1998; Corbit and Balleine, 2003; Killcross and Coutureau, 2003; Ostlund and Balleine, 2005). However, one difference between these areas emerges when the lesion operations are performed between the initial instrumental training and the devaluation test (rather than before training). In this case, dorsomedial striatal lesions continue to affect goal-directed behavior, but prelimbic lesions no longer affect it. These results suggest that both areas are involved in the *acquisition* of goal-directed behavior, but only dorsomedial striatum is implicated in its *expression* (Yin *et al.*, 2005).

In humans, there is now accumulating evidence to implicate the ventromedial prefrontal cortex in goal-directed learning (Balleine and O'Doherty, 2010). This area (and similarly positioned areas on the medial wall of prefrontal cortex in primates) is a likely homologue of the prelimbic cortex in rats, and also a key area in neuroeconomics due to many reports of correlates of expected value/utility there (see Chapters 8, 13, and 20 for more on this). Evidence tying this area to goal-directed behavior includes that value-related activity in this region tracks the current value of an instrumental action in a manner that mirrors goal-directed valuation. That is, following devaluation, activity decreases for an action associated with a devalued outcome relative to an action associated with a still valued action (Valentin *et al.*, 2007; de Wit *et al.*, 2009; Figure 21.4A,B). Furthermore, activity in this region also tracks measures related to the instrumental contingency (the causal relationship between an action and an outcome), sensitivity to which is another important feature of goal-directed control (Liljeholm *et al.*, 2011; Tanaka *et al.*, 2008). Also, studies examining whether expected value correlates in this region comply more with model-based (versus model-free) values, as predicted by computational models, have repeatedly shown evidence for model-based values there (Beierholm *et al.*, 2011; Daw *et al.*, 2011; Hampton *et al.*, 2006, 2008) (Figure 21.4C).

The human dorsomedial striatum has not been implicated as clearly in goal-directed behavior using

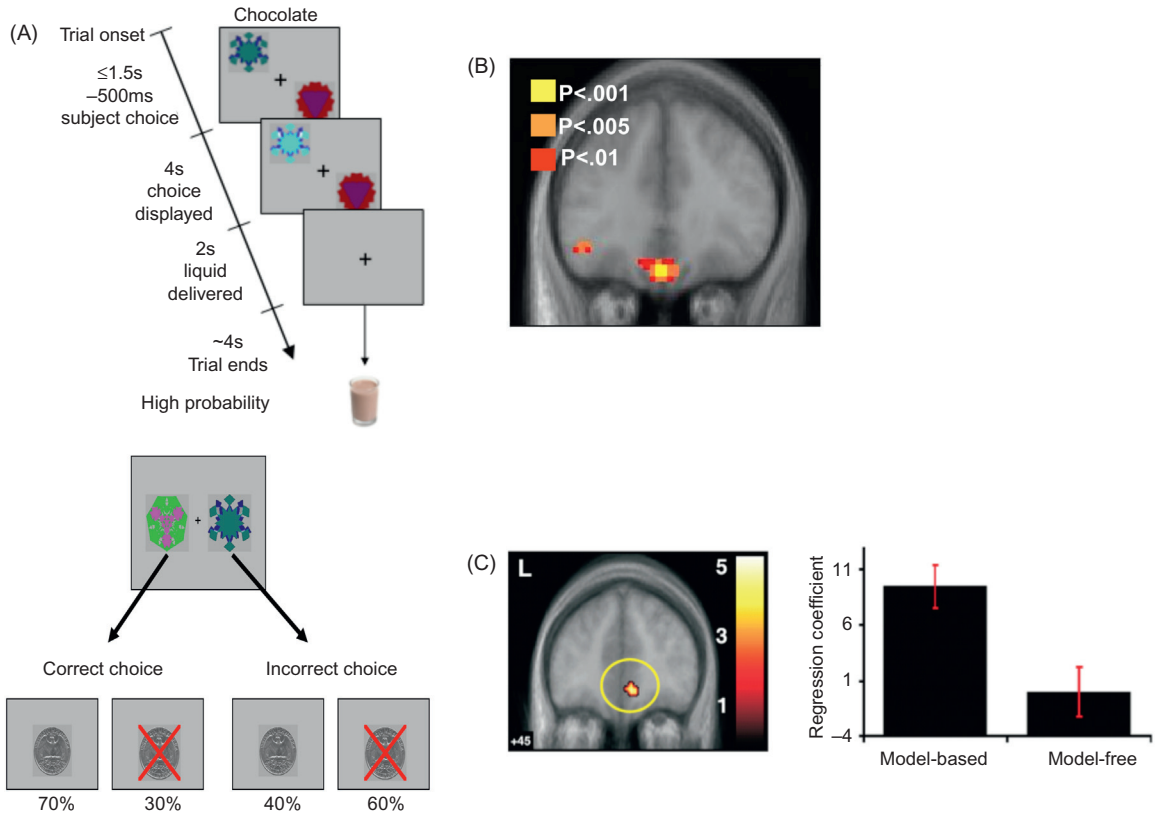

FIGURE 21.4 Human brain regions implicated in goal-directed control and in encoding model-based RL signals. (A) Instrumental-devaluation task used by Valentin and colleagues (2007) to uncover role for vmPFC in goal-directed control. Participants learned to choose between different actions denoted by discriminative stimuli which led to differing probabilities of obtaining a bolus of chocolate milk (illustrated here) or in a different condition, tomato juice. After training, one of the outcomes (either chocolate or tomato) were selectively devalued by feeding the participant to satiety on that outcome outside the scanner, and participants were placed back in the scanner and then invited to choose between the actions again. *Adapted with permission from Valentin* et al. *(2007).* (B) Task used in study by Hampton and colleagues (2006) to implicate vmPFC in encoding model-based valuations. Participants engaged in probabilistic reversal learning in which selection of one action denoted by a discriminative stimulus leads to a high probability of a monetary gain, while the other action leads to a high probability of monetary loss. After a period of time the contingencies reverse so that the previously rewarding action now predominantly yields losses while the previously punishing action yields gains. Hampton and colleagues (2006) constructed a computational model that incorporated knowledge of the reversal structure of the task and compared performance of that model against a model-free RL algorithm that did not incorporate knowledge of the task structure. (C) A region of vmPFC was found to correlate better with value signals generated by the model-based algorithm compared to the model-free algorithm, implicating this area in model-based valuation. *(B) and (C) adapted with permission from Hampton* et al. *(2006).*

devaluation protocols. However, contingency manipulations have implicated this area alongside the vmPFC (Liljeholm *et al.*, 2011; Tanaka *et al.*, 2008). Finally, the strength of the connection between ventromedial prefrontal cortex and dorsomedial striatum, as measured with DTI, has been shown to correlate with the degree of behavioral expression of goal-directed action selection across individuals (de Wit *et al.*, 2012).

The computational view of goal-directed behavior as supported by model-based RL raises several additional questions. First, Equation 21.1 shows how model-based valuation draws, separately, on two sorts of learning: state predictions (like the cognitive map), and learning about the current reward (incentive) value of particular goals or states. Moreover, combining these predictions to evaluate a candidate action involves a more active computation simulating future states. How does the brain implement these separate functions? Learning the relationships between states or stimuli, separate from reward value, is classically thought to implicate the hippocampus, especially in spatial tasks (Cohen and Eichenbaum, 1993; O'Keefe and Nadel, 1978). Tying this idea to model-based evaluation, Redish and colleagues have shown how representations of spatial location run ahead of rodents' current location at choice points in a spatial maze, entering different alternatives sequentially as though

simulating potential future trajectories (Johnson and Redish, 2007). Devaluation studies have failed to find an effect of hippocampal lesions on goal-directed lever pressing (Corbit *et al.*, 2002), but quite analogous studies involving an overtraining-dependent shift in behavior in spatial mazes do implicate hippocampus in the apparent spatial analogue of goal-directed action (Hartley and Burgess, 2005). This may suggest some suborganization of internal models by the type of task structure. An fMRI study in humans also implicates hippocampus in state prediction in a nonspatial associative learning task (Bornstein and Daw, 2012) and other fMRI studies also implicate additional areas of parietal and frontal cortices (Glascher *et al.*, 2010; Simon and Daw, 2011b). Finally, given the results discussed above, the circuit involving ventromedial PFC and dorsomedial striatum might be involved in tying these state predictions to incentive value.

Taken together, the above findings implicate specific segregated neural systems in each of the three types of learning outlined so far. However, given that each of these systems can exert control over behavior, this raises the question of how do these systems interact either competitively or cooperatively in order to mediate the control of actions?

WHAT IS THE NATURE OF THE INTERACTIONS BETWEEN THE DIFFERENT SYSTEMS?

Overall, given that all three of these systems can produce behaviors but these behaviors may often be mutually exclusive since there is only one body, the overriding sense of their interactions is competitive. For instance, we have already mentioned how the Pavlovian response to, for instance, salivate or approach in expectation of food can overcome the ability to learn, by instrumental means, to withhold these responses in order to obtain food. Similarly, we have argued that devaluation-insensitive habitual or devaluation-sensitive goal-directed behaviors dominate under different circumstances.

A major open and underexplored question, however, is how each system can come to control behavior at any one time. One possibility is that a separate arbitrator or controller sits on top of these systems and acts to gate their access to behavior. Alternatively, the systems might somehow competitively interact without the need for a separate arbitrator, e.g., through some sort of mutual inhibition at the point of action selection. Empirical evidence in favor of either of these possibilities is currently lacking. Moreover, although there have been suggestions of principles, already described, that may explain or rationalize the competition between goal-directed and habitual behaviors, these have yet to be mapped to any neural mechanism, nor have similar principles been elucidated for Pavlovian vs. instrumental competition.

Nevertheless, it is clear that all these systems do often compete to control action-selection in everyday life, with sometimes maladaptive consequences. In particular, habits can intrude to control performance under situations where a goal-directed action would be more appropriate. An example of this would be driving on the wrong side of the road while on vacation in a country with driving-side laws opposite to that in one's home country. On the computational analysis (Daw *et al.*, 2005; Keramati *et al.*, 2011; Simon and Daw, 2011a), although this might be a catastrophic error, this division of labor may still be justified on average by the computational savings (e.g., in time and energy) of adopting simpler model-free control.

The interaction between goal-directed and habitual systems, and particularly the situation where habits come to dominate behavior has become an area of considerable interest in neuropsychological models of addiction and other psychiatric disorders involving compulsive behaviors, such as obsessive compulsive disorder (Everitt and Robbins 2005; Redish *et al.*, 2008, 2012). Along these lines (Gillan *et al.*, 2011) it has recently been demonstrated that patients with obsessive compulsive disorder show a significant impairment in their ability to select goal-directed actions over habitual actions in a task in which the two systems were associated with competing actions, suggesting that these patients either have an over-active habitual system, an underactive goal-directed system or impaired arbitration.

Pavlovian and Instrumental Interactions

Although Pavlovian responses and instrumental actions also compete, with similarly deleterious consequences (Breland and Breland 1961; Dayan *et al.*, 2006), there are other classes of Pavlovian-instrumental interactions that are more facilitatory. In particular, there is a phenomenon known as *Pavlovian-to-instrumental transfer* (PIT) whereby ongoing instrumental action for reward (e.g., lever pressing) can be invigorated by the presentation of a Pavlovian stimulus that also predicts reward delivery. In rodents (Corbit *et al.*, 2001), this effect depends on the amygdala and ventral striatum (Corbit *et al.*, 2001); both areas have also been implicated in human studies (Bray *et al.*, 2008; Prevost *et al.*, 2012; Talmi *et al.*, 2008).

Although the literature on PIT has overwhelmingly focused on facilitatory interactions between Pavlovian and instrumental behaviors that are appetitively motivated, it is also the case that aversive Pavlovian cues can impact instrumental choice. Pavlovian cues predicting aversive outcomes such as electric shock reduce responding on an instrumental action for reward, a phenomenon known as conditioned suppression (Killcross *et al.*, 1997). Another example of where this type of adverse interaction might occur is in the phenomenon of "choking under pressure," whereby motor performance on a task under conditions of high stakes such as for a large monetary reward, breaks down relative to performance on the same task for a more modest incentive (Chib *et al.*, 2012; Mobbs *et al.*, 2009). In the study by Chib *et al.*, the degree of choking behavior exhibited by the subjects was found to be correlated with the level of deactivation in response to increasing incentives in the ventral striatum at the time of motor performance, implicating the ventral striatum in this type of adverse Pavlovian interaction alongside the facilitatory effects found in the appetitive domain.

A different sort of facilitatory Pavlovian-to-instrumental interaction is known as conditioned reinforcement. Here Pavlovian cues that are predictive of reward can act, like rewards, to drive new instrumental learning: e.g., a rat might learn to press a lever, which produces a light that was previously trained to predict food. (This is, at least notionally, different from PIT where the cues only facilitate previously learned responses, not learning itself.) Conditioned reinforcement is a laboratory model of the rewarding effects of something like money, which after all is only an intermediate means to obtain primary reward in the biological sense. Lesions suggest that the conditioned reinforcement effect also involves basolateral amygdala and ventral striatum – the same circuitry as PIT and Pavlovian learning more generally (Cador *et al.*, 1989). As discussed extensively in Chapter 15, this effect also has a clear resonance with model-free RL theories of instrumental learning and also with the responses of dopamine neurons, which can be driven by reward predictors as well as rewards. The relationship arises in computational strategies for model-free learning in settings like mazes or chess, where (unlike the simple one-step objective in Equation 21.1 here), multiple states and actions may intervene before an action is rewarded. In this case, the prediction error for model-free learning of future value (the "temporal-difference" error) incorporates information both from rewards and from predictors of future rewards, and either may train model-free action values $Q(s, a)$.

CONCLUSIONS: ALTERNATIVE SYSTEMS, MORE SYSTEMS

The idea that decisions are driven by multiple competing systems dates to the beginning of science: Plato likened the soul to a charioteer driving a team of winged horses, one honorable and guided by verbal commands, and another beastly and indecent (Plato, 1995: 428–437 B.C.). As we have mentioned already, numerous dual- or multiple-system frameworks have been proposed in cognitive psychology and behavioral economics. In our view, the multiple system approach detailed here is relatively unique when compared to these other (often more specialized) models in its long pedigree and breadth of support – comprising a century of behavioral, neural, and computational evidence in animals and humans. It also situates multiple system ideas in the context of other important concepts in neuroeconomics, such as theories concerning the role of dopamine in learning and the functions of value representations in the ventromedial prefrontal cortex. In general, a difficulty with dual-system views, including this one, is that differentiating their distinct contributions to behavior (which is, as we have stressed, often ambiguous) is at best laborious and at worst, assumption-bound. In both its psychological and computational incarnations, the framework described here is unusually specific in defining the contributions of the different controllers in ways that permit dissociating them experimentally. Moreover, the empirical relevance of these definitions is supported by the findings that they, indeed, produce clear neural and behavioral dissociations.

Although we would not go so far as to claim that other dual- or multiple-system theories are identical to this one, it does capture a number of common themes. Do these theories refer to the same systems we have delineated? To additional systems? Are there more controllers yet to be discovered? All these questions are difficult to answer. If the general view of the brain is that it consists of many interacting modules, then where to draw the boundaries between "systems" is somewhat of a matter of interpretation and nomenclature. For instance, much work in both human cognitive psychology and behavioral economics stresses a distinction between a controlled or deliberative mode, and an automatic behavioral mode for the monitoring and selection of behavior (Kahneman, 2003; Norman and Shallice, 1986; Schneider and Shiffrin, 1977). This corresponds plausibly to the distinction put forward here, with both habitual and Pavlovian behaviors as distinct types of automatic control. However, dual process theories from cognitive psychology tend to associate the deliberative mode with linguistic (or sometimes

rule-based or explicit) reasoning, which clearly does not contribute to goal-directed rat lever pressing. Nonetheless, such reasoning falls under the broader purview of model-based learning, and might be considered part of an augmented or evolved version of the rodent's goal-directed controller.

Similarly, we have here suggested that the hippocampus may serve as part of a model-based system, particularly contributing to internal models for certain tasks such as spatial navigation. This fits well with another influential multiple-systems hypothesis arising from the memory literature, which distinguishes a declarative memory system centered on the hippocampus from a subcortical system for procedural memories much like habitual and Pavlovian learning (Squire, 1992). While this theory tends to emphasize the content of different memories and the degree of conscious (explicit) access to those memories as opposed to the role of these systems in guiding behavior, the basic division corresponds reasonably well with our framework. On the other hand, combining these ideas with the same three-system framework we present, Lengyel and Dayan (2007) instead suggest that hippocampus subserves a distinct, fourth system for *episodic* control, separate from the goal-directed controller. As with model-based RL, this system is envisioned to compute values using the right hand side of Equation 21.1, but via a different approximation to that average than was employed in previous theories, computing values as a function of particular previous episodes rather than their statistical summary in an internal model.

Whether this (or for that matter, linguistic or rule-based control) is best understood as a refinement in the characterization of a single model-based controller, versus a truly dissociable system and mode of computation, might best be judged by whether clear double dissociations can be found between them and traditional goal-directed control. This is the same manner that goal-directed and habitual systems have been dissociated. No similar evidence yet exists dissociating goal-directed from a putative episodic (or linguistic) controller.

Finally, perhaps the most important aspect of multiple system theories in behavioral and neuroeconomics has been their implications for errors and problems of self-control. Particularly influential in this respect has been an emphasis, pervasive in some theories going back even to Plato's horses, on a distinction between a "cold," rational, logical system for guiding behavior, and a "hot", affectively charged or impulsive system (Damasio, 1994; Loewenstein, 1996; Weber and Johnson, 2009). Various deviations from normative choice, such as framing effects or hyperbolic time discounting, have been argued to arise from "hot" interference (Loewenstein, 1996). In intertemporal choice, for instance, a "patient" system that computes values with a slow discount rate is suggested to compete with a more "impatient" system that computes values with a high discount rate, with these influences superimposing to produce quasi-hyperbolic discounting (Laibson, 1997; Thaler and An, 1981). The jury is still out as to whether such a putative distinction between inter-temporal processes is implemented at the neural level, with initial claims in this regard (McClure *et al.*, 2004) having been subsequently challenged by evidence in favor of an integrated (single process) account (Kable and Glimcher, 2007). Nevertheless, it is clearly the case that multiple-system theories have rich implications for self-control and rationality. How might these theories relate to the multiple-learning system view described here?

In both our phylogenetic and computational presentations, we have stressed the differences between the systems in terms of the mechanisms or computational techniques they bring to bear in serving common and adaptive ends of obtaining reward or avoiding punishment. In this respect, our characterization eschews any distinction between systems as more or less "rational" or "emotional." That said, the behavioral repertoire that the Pavlovian system uses to pursue adaptive ends centers, in the appetitive domain, mainly on reflexes for approaching and consuming rewards. These could certainly be interpreted to have a "hot" or "impulsive" character, as could the arousing and facilitating influences of Pavlovian learning on instrumental behaviors (PIT, discussed above).

Thus, we have suggested (Dayan *et al.*, 2006) that the Pavlovian system in particular captures (and reinterprets) many of the "hot" or "impulsive" influences described by other authors. Such influences are clearly applicable to many tests of self-control in the face of direct temptation (e.g., the famous "marshmallow test" of Walter Mischel and his colleagues; Mischel, 1974), though less obviously to choices in more descriptive settings. Another important feature of Pavlovian learning is that in many preparations, the strength of conditioning is strongly dependent on the temporal interval between the conditioned and unconditioned stimulus (Holland, 1980), such that (at least, when other time intervals in the experiments are held fixed) Pavlovian conditioned responses are strongest for a shorter interval between stimulus and outcome, and become progressively weaker as the interval increases. This feature of temporal dependence in Pavlovian conditioning may produce some effects resembling short-horizon discounting. Finally, in fMRI studies, the same brain systems known to be core parts of the Pavlovian system and to mediate Pavlovian-instrumental interactions, namely the ventral striatum and the amygdala, are most often identified as being active under

situations where an individual's choices are prone to irrational biases such as framing or endowment effects, (De Martino *et al.*, 2006, 2009).

In all, the three-system approach we have described captures many themes and features of other frameworks in this area. In particular, interpreting issues of impulsivity and self-control in terms of Pavlovian influences casts them in a different and potentially revealing light. For the same reasons, it also points to new questions. For instance, as we have mentioned, there has been considerable focus on understanding the competitive interactions between habitual and goal-directed instrumental control, which are also important to self-control issues in overcoming maladaptive habits, as in drug abuse. But to the extent that self-control and related problems at the heart of neuroeconomics instead center on overcoming Pavlovian responses, these are theoretically and empirically more underexplored. For instance, under what circumstances can Pavlovian influences be overcome at all? Can these interactions be understood in terms of rational cost–benefit tradeoffs? Although Pavlov's original work is a century old, implications like these remain to be explored.

References

Adams, C., 1982. Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q. J. Exp. Psychol. Sect. B. 34 (2), 77–98.

Aggarwal, M., Hyland, B.I., Wickens, J.R., 2012. Neural control of dopamine neurotransmission: implications for reinforcement learning. Eur. J. Neurosci. 35 (7), 1115–1123.

Alexander, G.E., Crutcher, M.D., 1990. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci. 13 (7), 266–271.

Balleine, B.W., Daw, N.D., O'Doherty, J.P., 2008. Multiple forms of value learning and the function of dopamine. In: Glimcher, P.W., Camerer, C.F., Poldrack, R.A., Fehr, E. (Eds.), Neuroeconomics: Decision Making and the Brain. Academic Press, New York.

Balleine, B.W., Dickinson, A., 1998. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 37 (4–5), 407–419.

Balleine, B.W., O'Doherty, J.P., 2010. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 35 (1), 48–69.

Barto, A.G., 1995. Adaptive critics and the basal ganglia. In: Houk, J. C., Davis, J.L., Beiser, D.G. (Eds.), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 215–232.

Beierholm, U.R., Anen, C., Quartz, S., Bossaerts, P., 2011. Separate encoding of model-based and model-free valuations in the human brain. NeuroImage. 58 (3), 955–962.

Berg, H.C., Brown, D.A., 1972. Chemotaxis in escherichia coli analysed by three-dimensional tracking. Nature. 239 (5374), 500–504.

Bindra, D., 1976. A Theory of Intelligent Behavior. Wiley-Interscience, Oxford, England.

Bornstein, A.M., Daw, N.D., 2012. Dissociating hippocampal and striatal contributions to sequential prediction learning. European J. Neurosci. 35 (7), 1011–1023.

Bray, S., Rangel, A., Shimojo, S., Balleine, B., O'Doherty, J.P., 2008. The neural mechanisms underlying the influence of Pavlovian cues on human decision making. J. Neurosci. 28 (22), 5861–5866.

Breland, K., Breland, M., 1961. The misbehavior of organisms. Am. psychol. 16 (11), 681.

Brembs, B., Lorenzetti, F.D., Reyes, F.D., Baxter, D.A., Byrne, J.H., 2002. Operant reward learning in Aplysia: neuronal correlates and mechanisms. Science. 296 (5573), 1706–1709.

Cador, M., Robbins, T., Everitt, B., 1989. Involvement of the amygdala in stimulus–reward associations: interaction with the ventral striatum. Neuroscience. 30 (1), 77–86.

Chib, V.S., De Martino, B., Shimojo, S., O'Doherty, J.P., 2012. Neural mechanisms underlying paradoxical performance for monetary incentives are driven by loss aversion. Neuron. 74 (3), 582–594.

Cohen, N.J., Eichenbaum, H., 1993. Memory, amnesia, and the hippocampal system. MIT press, Cambridge, MA.

Cook, D.G., Carew, T.J., 1986. Operant conditioning of head waving in Aplysia. Proc. Natl. Acad. Sci. U.S.A. 83 (4), 1120–1124.

Corbit, L.H., Balleine, B.W., 2003. The role of prelimbic cortex in instrumental conditioning. Behav. Brain Res. 146 (1–2), 145–157.

Corbit, L.H., Muir, J.L., Balleine, B.W., 2001. The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. J. Neurosci. 21 (9), 3251–3260.

Corbit, L.H., Ostlund, S.B., Balleine, B.W., 2002. Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. J. Neurosci. 22 (24), 10976–10984.

Damasio, A.R., 1994. Descartes' Error: Emotion, Rationality and the Human Brain. Putnam, New York.

Davey, G.C., 1992. Classical conditioning and the acquisition of human fears and phobias: A review and synthesis of the literature. Adv. Behav. Res. Ther. 14 (1), 29–66.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron. 69 (6), 1204–1215.

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8 (12), 1704–1711.

Dayan, P., Niv, Y., Seymour, B., Daw, N.D., 2006. The misbehavior of value and the discipline of the will. Neural Netw. 19 (8), 1153–1160.

De Martino, B., Kumaran, D., Holt, B., Dolan, R.J., 2009. The neurobiology of reference-dependent value computation. J. Neurosci. 29 (12), 3833–3842.

De Martino, B., Kumaran, D., Seymour, B., Dolan, R.J., 2006. Frames, biases, and rational decision-making in the human brain. Science. 313 (5787), 684–687.

de Wit, S., Corlett, P.R., Aitken, M.R., Dickinson, A., Fletcher, P.C., 2009. Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. J. Neurosci. 29 (36), 11330–11338.

de Wit, S., Watson, P., Harsay, H.A., Cohen, M.X., van de Vijver, I., Ridderinkhof, K.R., 2012. Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. J. Neurosci. 32 (35), 12066–12075.

Dickinson, A., 1985. Actions and habits: the development of a behavioural autonomy. Philos. Trans. R. Soc. Lond. B Biol. Sci. 308, 67–78.

Dickinson, A., 1996. Bidirectional instrumental conditioning. Q. J. Exp. Psychol. Sect. B. 49 (4), 289–306.

Everitt, B.J., Robbins, T.W., 2005. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat. Neurosci. 8 (11), 1481–1489.

Faure, A., Haberland, U., Conde, F., El Massioui, N., 2005. Lesion to the nigrostriatal dopamine system disrupts stimulus–response habit formation. J. Neurosci. 25, 2771–2780.
Freud, S., 1961. The Ego and the Id, vol. 19. Hogarth Press, London.
Gillan, C.M., Papmeyer, M., Morein-Zamir, S., et al., 2011. Disruption in the balance between goal-directed behavior and habit learning in obsessive–compulsive disorder. Am. J. Psychiatry. 168 (7), 718–726.
Glascher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 66 (4), 585–595.
Gottfried, J.A., O'Doherty, J., Dolan, R.J., 2002. Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. J. Neurosci. 22 (24), 10829–10837.
Gottfried, J.A., O'Doherty, J., Dolan, R.J., 2003. Encoding predictive reward value in human amygdala and orbitofrontal cortex. Science. 301 (5636), 1104–1107.
Groenewegen, H.J., Russchen, F.T., 1984. Organization of the efferent projections of the nucleus accumbens to pallidal, hypothalamic, and mesencephalic structures: a tracing and immunohistochemical study in the cat. J. Comp. Neurol. 223 (3), 347–367.
Hampton, A.N., Bossaerts, P., O'Doherty, J.P., 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. 26 (32), 8360–8367.
Hampton, A.N., Bossaerts, P., O'Doherty, J.P., 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. U.S.A. 105 (18), 6741–6746.
Hartley, T., Burgess, N., 2005. Complementary memory systems: competition, cooperation and compensation. Trends Neurosci. 28 (4), 169–170.
Hatfield, T., Han, J.S., Conley, M., Gallagher, M., Holland, P., 1996. Neurotoxic lesions of basolateral, but not central, amygdala interfere with Pavlovian second-order conditioning and reinforcer devaluation effects. J. Neurosci. 16 (16), 5256–5265.
Hershberger, W.A., 1986. An approach through the looking glass. Learn. Behav. 14 (4), 443–451.
Holland, P.C., 1980. CS–US interval as a determinant of the form of Pavlovian appetitive conditioned responses. J. Exp. Psychol. Anim. Behav. Process. 6 (2), 155–174.
Holland, P.C., 2004. Relations between Pavlovian-instrumental transfer and reinforcer devaluation. J. Exp. Psychol. Animal Behav. Proc. 30, 104–117.
Hull, C., 1943. Principles of Behavior. Appleton-Century-Crofts, New York, NY.
James, W., 1950. The Principles of Psychology, vol. 1. Dover, New York.
Johnson, A., Redish, A.D., 2007. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. 27 (45), 12176–12189.
Jueptner, M., Frith, C.D., Brooks, D.J., Frackowiak, R.S., Passingham, R.E., 1997. Anatomy of motor learning. II. Subcortical structures and learning by trial and error. J. Neurophysiol. 77, 1325–1337.
Kable, J.W., Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. Nat. Neurosci. 10 (12), 1625–1633.
Kahneman, D., 2003. A perspective on judgment and choice. Am. Psychol. 58, 697–720.
Keramati, M., Dezfouli, A., Piray, P., 2011. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput. Biol. 7 (5), e1002055.
Killcross, S., Coutureau, E., 2003. Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb. Cortex. 13 (4), 400–408.
Killcross, S., Robbins, T.W., Everitt, B.J., 1997. Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala. Nature. 388 (6640), 377–380.
Konorski, J., 1948. Conditioned Reflexes and Neuron Organization. Cambridge University Press, Cambridge.
Laibson, D., 1997. Golden eggs and hyperbolic discounting. Q. J. Econ. 112 (2), 443–477.
Lehéricy, S., Benali, H., Van de Moortele, P.F., Pélégrini-Issac, M., Waechter, T., Ugurbil, K., 2005. Distinct basal ganglia territories are engaged in early and advanced motor sequence learning. PNAS. 102, 12566–12571.
Lengyel, M., Dayan, P., 2007. Hippocampal contributions to control: the third way. Adv. Neural Inf. Process. Syst. 20, 889–896.
Liljeholm, M., Tricomi, E., O'Doherty, J.P., Balleine, B.W., 2011. Neural correlates of instrumental contingency learning: differential effects of action–reward conjunction and disjunction. J. Neurosci. 31 (7), 2474–2480.
Loewenstein, G., 1996. Out of control: visceral influences on behavior. Organ. Behav. Hum. Decis. Process. 65 (3), 272–292.
Loewenstein G., O'Donoghue T., 2004. Animal spirits: affective and deliberative processes in economic behavior. Working Paper 04–14, Center for Analytic Economics, Cornell University.
McClure, S.M., Laibson, D.I., Loewenstein, G., Cohen, J.D., 2004. Separate neural systems value immediate and delayed monetary rewards. Science. 306 (5695), 503–507.
Mischel, W., 1974. Processes in delay of gratification. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 7. Academic Press, New York.
Mobbs, D., Hassabis, D., Seymour, B., et al., 2009. Choking on the money: reward-based performance decrements are associated with midbrain activity. Psychol. Sci. 20 (8), 955–962.
Montague, P.R., Dayan, P., Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16 (5), 1936–1947.
Morrison, S.E., Salzman, C.D., 2011. Representations of appetitive and aversive information in the primate orbitofrontal cortex. Ann. N. Y. Acad. Sci. 1239, 59–70.
Mowrer, O.H., 1947. On the dual nature of learning—a re-interpretation of "conditioning" and "problem-solving. Harv. Educ. Rev. 17, 102–148.
Norman, D.A., Shallice, T., 1986. Attention to action: Willed and Automatic control of behaviour. In: Schwartz, G.E., Shapiro, D. (Eds.), Consciousness and Self-regulation. Plenum Press, New York, NY.
O'Doherty, J., Deichmann, R., Critchley, H.D., Dolan, R.J., 2002. Neural responses during anticipation of a primary taste reward. Neuron. 33 (5), 815–826.
O'Keefe, J., Nadel, L., 1978. The Hippocampus as a Cognitive Map. Clarendon, Oxford.
Ostlund, S.B., Balleine, B.W., 2005. Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. J. Neurosci. 25 (34), 7763–7770.
Ostlund, S.B., Balleine, B.W., 2007. Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. J. Neurosci. 27 (18), 4819–4825.
Pare, D., Quirk, G.J., Ledoux, J.E., 2004. New vistas on amygdala networks in conditioned fear. J. Neurophysiol. 92 (1), 1–9.
Parkinson, J., Dalley, J., Cardinal, R., et al., 2002. Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. Behav. Brain Res. 137 (1–2), 149–163.
Parkinson, J.A., Olmstead, M.C., Burns, L.H., Robbins, T.W., Everitt, B.J., 1999. Dissociation in effects of lesions of the nucleus accumbens core and shell on appetitive Pavlovian approach behavior

and the potentiation of conditioned reinforcement and locomotor activity by D-amphetamine. J. Neurosci. 19 (6), 2401–2411.
Paton, J.J., Belova, M.A., Morrison, S.E., Salzman, C.D., 2006. The primate amygdala represents the positive and negative value of visual stimuli during learning. Nature. 439 (7078), 865–870.
Pavlov, I.P., 1927. Conditioned Reflexes. Oxford University Press, Oxford.
Plato, 1995. Phaedrus. Hackett Publishing Company, Indianapolis, Indiana, ca. 428–347 B.C.
Prevost, C., Liljeholm, M., Tyszka, J.M., O'Doherty, J.P., 2012. Neural correlates of specific and general Pavlovian-to-Instrumental Transfer within human amygdalar subregions: a high-resolution fMRI study. J. Neurosci. 32 (24), 8383–8390.
Price, J.L., Amaral, D.G., 1981. An autoradiographic study of the projections of the central nucleus of the monkey amygdala. J. Neurosci. 1 (11), 1242–1259.
Redish, A.D., Jensen, S., Johnson, A., 2008. A unified framework for addiction: vulnerabilities in the decision process. Behav. Brain Sci. 31 (4), 415–437 (discussion 437–487).
Reynolds, J.N., Wickens, J.R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. 15 (4–6), 507–521.
Schacter, D., Addis, D., Buckner, R., 2007. Remembering the past to imagine the future: the prospective brain. Nat. Rev. Neurosci. 8 (9), 657–661.
Schneider, W., Shiffrin, R.M., 1977. Controlled and automatic human information processing: I. Detection, search, and attention. Psychol. Rev. 84 (1), 1–66.
Schoenbaum, G., Chiba, A.A., Gallagher, M., 1998. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. Nat. Neurosci. 1 (2), 155–159.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science. 275 (5306), 1593–1599.
Sheffield, F.D., 1965. Relation betwen classical and instrumental conditioning. In: Prokasy, W.F. (Ed.), Classical Conditioning. Appleton–Century-Crofts, New York, NY, pp. 302–322.
Sherrington, C., 1906. The Integrative Action of the Nervous System. Yale University Press, New Haven.
Simon, D.A., Daw, N.D., 2011a. Environmental statistics and the trade-off between model-based and TD learning in humans. Adv. Neural Inf. Process. Syst. 24.
Simon, D.A., Daw, N.D., 2011b. Neural correlates of forward planning in a spatial decision task in humans. J. Neurosci. 31 (14), 5526–5539.
Skinner, B.F., 1962. Two "synthetic social relations". J. Exp. Anal. Behav. 5 (4), 531–533.
Squire, L.R., 1992. Declarative and nondeclarative memory: multiple brain systems supporting learning and memory. J. Cogn. Neurosci. 4 (3), 232–243.
Sutton, R.S., 1988. Learning to predict by the methods of temporal differences. Mach. Learn. 3 (1), 9–44.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press.
Talmi, D., Seymour, B., Dayan, P., Dolan, R.J., 2008. Human Pavlovian-instrumental transfer. J. Neurosci. 28 (2), 360–368.
Tanaka, S.C., Balleine, B.W., O'Doherty, J.P., 2008. Calculating consequences: brain systems that encode the causal effects of actions. J. Neurosci. 28 (26), 6750–6755.
Thaler, R.H., An, H.M.S., 1981. Economic Theory of Self Control. J. Polit. Econ. 89 (2), 392–406.
Thorndike, E.L., 1898. Animal Intelligence: An Experimental Study of the Associative Processes in Animals. Vol Monograph Supplements, No. 8. Macmillan, New York.
Tolman, E., 1948. Cognitive maps in rats and men. Psychol. Rev. 55 (4), 189–208.
Tricomi, E., Balleine, B.W., O'Doherty, J.P., 2009. A specific role for posterior dorsolateral striatum in human habit learning. Eur. J. Neurosci. 29, 2225–2232.
Tully, T., Quinn, W.G., 1985. Classical conditioning and retention in normal and mutant Drosophila melanogaster. J. Comp. Physiol. A. 157 (2), 263–277.
Valentin, V.V., Dickinson, A., O'Doherty, J.P., 2007. Determining the neural substrates of goal-directed learning in the human brain. J. Neurosci. 27 (15), 4019–4026.
Walters, E.T., Carew, T.J., Kandel, E.R., 1981. Associative learning in aplysia: evidence for conditioned fear in an invertebrate. Science. 211 (4481), 504–506.
Wang, L.P., Li, F., Wang, D., Xie,K., Shen, X., Tsien, J.Z., 2011. NMDA receptors in dopaminergic neurons are crucial for habit learning. *Neuron*. 72 (6), 1055–1066.
Weber, E.U., Johnson, E.J., 2009. Mindful judgment and decision making. Annu. Rev. Psychol. 60, 53–85.
Wunderlich, K., Dayan, P., Dolan, R.J., 2012. Mapping value based planning and extensively trained choice in the human brain. Nat. Neurosci. 15 (5), 786–791.
Yin, HH, Knowlton, BJ, Balleine, BW., 2004. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur. J. Neurosci. 19 (1), 181–189.
Yin, H.H., Ostlund, S.B., Knowlton, B.J., Balleine, B.W., 2005. The role of the dorsomedial striatum in instrumental conditioning. Eur. J. Neurosci. 22 (2), 513–523.