Published in final edited form as:

Annu Rev Psychol. 2017 January 03; 68: 73–100. doi:10.1146/annurev-psych-010416-044216.

Learning, Reward, and Decision Making

John P. O'Doherty, **Jeffrey Cockburn**[#], and **Wolfgang M. Pauli**[#]

Division of Humanities and Social Sciences and Computation and Neural Systems Program, California Institute of Technology, Pasadena, California 91125; jdoherty@caltech.edu

[#] These authors contributed equally to this work.

Abstract

In this review, we summarize findings supporting the existence of multiple behavioral strategies for controlling reward-related behavior, including a dichotomy between the goal-directed or model-based system and the habitual or model-free system in the domain of instrumental conditioning and a similar dichotomy in the realm of Pavlovian conditioning. We evaluate evidence from neuroscience supporting the existence of at least partly distinct neuronal substrates contributing to the key computations necessary for the function of these different control systems. We consider the nature of the interactions between these systems and show how these interactions can lead to either adaptive or maladaptive behavioral outcomes. We then review evidence that an additional system guides inference concerning the hidden states of other agents, such as their beliefs, preferences, and intentions, in a social context. We also describe emerging evidence for an arbitration mechanism between model-based and model-free reinforcement learning, placing such a mechanism within the broader context of the hierarchical control of behavior.

Keywords

model based; model free; instrumental; Pavlovian; cognitive map; outcome valuation

INTRODUCTION

All organisms, including humans, face the fundamental challenge of the need to interact effectively with the environment in a manner that maximizes the prospects of obtaining the resources needed to survive and procreate while minimizing the prospect of encountering situations leading to harm. Organisms have evolved a variety of strategies to solve this problem. Accumulating evidence suggests that these distinct strategies coexist in the human brain. In this review, we outline evidence for the existence of these multiple systems of behavioral control and describe how they can be either interdependent or mutually interfering depending on the situation. We establish the role that predictions play in guiding these different behavioral systems and consider how these systems differ in the ways in which they develop their predictions. Finally, we evaluate the possibility that an additional

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

system, used for performing learning and inference in social contexts, is present in the human brain.

Multiple Strategies for Behavioral Control

Perhaps one of the most fruitful questions that may be answered by an understanding of the brain's varied control strategies is whether behavior is motivated by the onset of a stimulus or is directed toward a goal outcome. Historically, habitual responses that are elicited by the perception of a stimulus regardless of the action's consequences (Thorndike 1898) have been contrasted with goal-directed actions that are deliberatively dispatched to achieve a goal (Tolman 1948). Theory and evidence have resolved arguments as to whether human (and animal) behavior is ruled by one strategy or the other by suggesting that both types of behavioral control coexist. In the following sections, we outline some of the behavioral evidence in support of multiple strategies for behavioral control.

Stimulus-Driven Control

Stimulus-driven control refers to a class of behaviors that are expressed in response to the onset of an unanticipated external stimulus. Because these behaviors are instigated by a particular stimulus or class of stimuli, they are cognitively efficient, automatic, and rapidly deployed. However, because they are initiated without consideration of the organism's goals or subsequent outcomes, stimulus-driven behaviors can suffer from being overly rigid, especially in a volatile environment.

Reflexes are perhaps the most primitive form of adaptive response to environmental challenges. Reflexes are stereotyped in that sensory stimuli have innate (unlearned) activating tendencies; thus, reflexes do not depend on synaptic plasticity and are often implemented at the level of the spinal cord and brainstem (Thibodeau & Patton 1992). Reflexes have a long evolutionary history because they are present in organisms from the simplest, such as bacteria, to the most complex, such as humans, and because analogous motor reflexes to the same stimulus are present across species. Examples of reflexes include the withdrawal reflex that comes from touching a hot surface, the startle response that is elicited in response to sudden stimuli, and the salivatory response to the presentation of food. Reflexes are considered advantageous. For example, the withdrawal reflex helps to avoid tissue damage, the startle response facilitates successful escape responses, and the salivary response aids in the consumption and digestion of food.

Reflexes are fundamentally reactive in that an unanticipated triggering stimulus elicits a preprogrammed response. However, being able to issue responses in a prospective manner, in anticipation of an event that requires a response, provides significant advantages. For example, digestion can be aided by producing saliva prior to the arrival of food, and personal harm may be avoided by steering clear of a hot surface without having to reflexively retreat from it. Pavlovian conditioning, also referred to as classical conditioning, is a means by which an organism can learn to make predictions about the subsequent onset of behaviorally significant events and leverage these predictions to initiate appropriate anticipatory behaviors (Pavlov 1927). As is the case for reflexes, Pavlovian learning is present in many

invertebrates, including insects such as *Drosophila* (Tully & Quinn 1985) and even sea slugs (*Aplysia*; Walters et al. 1981), but also in vertebrates, including humans (Davey 1992).

The type of behavior emitted in response to the stimulus depends on the form of outcome the stimulus is paired with (Jenkins & Moore 1973). For instance, a cue paired with the subsequent delivery of food will result in the acquisition of a salivary response, whereas a cue paired with aversive thermal heat will elicit avoidance behavior. Different classes of Pavlovian conditioned responses have been identified. Some are almost identical to the unconditioned responses elicited by the stimuli that trigger them, but other conditioned Pavlovian responses are more distinct. For example, in addition to salivating in response to a food predictive cue, animals also typically orient toward the site of anticipated food delivery (Konorski & Miller 1937).

Although the adaptive advantages of anticipatory behavior are clear, Pavlovian learning is limited to learning about events that occur independent of the organism's behavior. In other words, Pavlovian learning may help an organism prepare for the arrival of food, but it won't help that organism procure its next meal. To increase the possibility of being able to actively attain rewards, many organisms are also equipped with instrumental conditioning, a mechanism that allows them to learn to perform specific yet arbitrary behavioral responses (such as a lever press) in a specific context. In the simplest form of instrumental conditioning, specific stimulus–response patterns are acquired by virtue of the extent to which a particular response gives rise to positive (i.e., the receipt of a reward) or negative (i.e., avoidance of an aversive outcome) reinforcement. This strategy provides significant benefits in terms of cognitive efficiency, speed, and accuracy; however, these benefits come at a cost. Critically, the execution of this class of behavior does not involve an anticipation of a particular outcome (Thorndike 1898); thus, behavior can become habitual, making it difficult to flexibly adjust the behavior should outcome valuation suddenly change. Thus, to the organism's potential detriment, habits may persist even if their outcomes are no longer beneficial. This persistence is suggested to give rise to various forms of addiction (Everitt & Robbins 2016).

Goal-Directed Control

Goal-directed control refers to a class of instrumental behaviors that appear to be motivated by and directed toward a specific outcome. Whereas stimulus-driven control can be thought of as retrospective in that it depends on integrating past experience, goal-directed control may be thought of as prospective in that it leverages a cognitive map of the decision problem to flexibly revalue states and actions (Tolman 1948). Leveraging this map in conjunction with the organism's internal goals facilitates a highly flexible control system, allowing the organism to adapt to changes in the environment without having to resample environmental contingencies directly. However, the necessity of interrogating a cognitive map in order to generate a behavioral plan makes goal-directed control cognitively demanding and slow.

Goal-directed control has been experimentally distinguished from habitual behavior in a study involving training an animal to perform unique actions (e.g., pressing a lever or pulling a chain) in order to obtain unique food outcomes, then devaluing one of the outcomes by pairing it with illness (Balleine & Dickinson 1991). If the animal is behaving in

a goal-directed manner, it should be less likely to elicit the action that had been associated with the now-devalued outcome. Indeed, some animals (Dickinson 1985) and humans (Valentin et al. 2007) have been shown to exhibit goal-directed control.

Evidence for the Coexistence of Multiple Control Systems

Although Dickinson & Balleine (1994) demonstrated that rats are capable of performing in a goal-directed manner, Dickinson et al. (1995) also showed that those same animals may also exhibit habitual tendencies. For example, after animals were exposed to extensive training, they were found to persistently elicit responses associated with devalued outcomes (Dickinson et al. 1983). These findings led to the proposal that animals were no longer sensitive to the value of the outcome, but that their behavior was instead driven by the stimulus that had been paired with response. Thus, reward schedules and degree of experience guide, at least in part, the control strategy deployed by the animal. Dickinson et al. (1983) concluded that both habitual and goal-directed systems of control are present in rodents and that these two systems manifest themselves in behavior under different circumstances. Using a similar overtraining manipulation to that performed in rodents, Tricomi et al. (2009) showed that humans also exhibit reduced outcome sensitivity consistent with the behavioral expression of habit.

Even though the distinction between habitual and goal-directed control is often conceptualized and investigated within the context of instrumental behavior, there is tentative evidence that a similar distinction can be made for Pavlovian behavior. Critically, the core criterion to distinguish habitual from goal-directed behavior in the instrumental domain is also present for conditioned Pavlovian responses: Some Pavlovian responses are more sensitive (Dayan & Berridge 2014) than others to outcome value (Nasser et al. 2015). Nevertheless, Pavlovian conditioned responses are often considered to be habitual in a manner analogous to habits in the instrumental domain; this conception of Pavlovian responses gives rise to the prevalent assumption that incremental synaptic plasticity implements the acquisition of Pavlovian contingencies (Rescorla & Wagner 1972). However, this form of habitual Pavlovian conditioning cannot account for findings showing altered patterns in the conditioned response immediately after devaluation and prior to any resampling of the environment's contingencies (Dayan & Berridge 2014). Despite the evidence for the existence of distinct habitual and goal-directed strategies within Pavlovian learning, the majority of the research on multiple control systems has been performed using instrumental conditioning; we also focus on instrumental conditioning in the remainder of this review, although we revisit the Pavlovian case in the section Model-Free and Model-Based Pavlovian Learning.

Why Multiple Systems?

Given that all of the different strategies for controlling behavior that we have described, from reflexes to goal-directed behavior, seem to be present in humans, a natural question follows: Why have all of these systems continued to coexist simultaneously? In other words, why are humans still endowed with the capacity for less flexible Pavlovian reflexes when they have machinery enabling more flexible goal-directed actions instead? One explanation could be that these behavioral control systems coexist because evolutionary adaptation

occurred incrementally. The adaptations allowing goal-directed actions may simply have occurred through the addition of new brain circuitry without the refurbishment or repurposing of control systems already in place, similar to adding a modern extension to an older building. However, this seems unlikely given the inefficiencies (both biologically and functionally) associated with adopting a multicontroller strategy in the absence of some additional benefit.

A second, more compelling possible explanation for the coexistence of multiple behavioral control systems is that the brain's control systems share mutually beneficial interdependencies. Evolutionarily recent regions may depend on the computations performed by more primal regions. Primal regions may also take advantage of the experience that comes with more complex control strategies, as well as more evolutionarily recently developed brain regions, which afford powerful domain-general computational functions to existing decision-making strategies. In other words, primal control systems could offer the scaffolding required for more advanced control systems, and the strategic guidance of advanced systems could help primal systems build adaptive associations more efficiently. Indeed, theoretical work (Sutton 1990) has demonstrated that stimulus-driven learning can be significantly improved when guided by a goal-directed system, and experimental work suggests that these interactions take place in the human brain (Doll et al. 2011).

Yet another benefit of multiple behavioral control systems is rooted in the mutually exclusive challenges faced by most organisms. Each system offers a different solution for the trade-off between accuracy, speed, experience, and (computational) efficiency. Goal-directed control typically moves an organism toward goal satisfaction more reliably than other systems, but its flexibility is cognitively demanding and deployment is relatively slow. A goal-directed strategy could offer significant advantages to a predator stalking its prey but prove ruinous for the prey when a swift retreat is required. Conversely, although stimulus-driven behaviors may not always meet an organism's current needs, particularly in a volatile environment, they can be deployed quickly and require less computational resources because they rely on simple stimulus–response associations rather than a rich cognitive map.

The environment presents complex challenges to survival, the range of which demand mutually exclusive strategies to tackle them in an adaptive manner. Organisms stand to gain the best of all worlds by preserving and adaptively deploying multiple control strategies that meet these challenges. However, before we can begin to understand how the brain handles the coexistence of these different forms of behavior, we first need to consider computational theories of value-based decision making, learning, and action selection to fully grasp the nature of the computations implemented in partially separable networks of brain areas.

ALGORITHMS FOR LEARNING AND DECISION MAKING

A central notion in most (e.g., Balleine et al. 2009, Camerer et al. 2005, Glimcher et al. 2013, Padoa-Schioppa & Assad 2006, Platt & Glimcher 1999, Rangel et al. 2008) but not all (see Gigerenzer & Gaissmaier 2011, Strait et al. 2014) theories of value-based decision making as applied to the brain is that, to establish which option to take, an agent must first

compute a representation of the expected value or utility that will follow from selecting a particular option. This computation facilitates a comparative process, allowing the agent to identify and pursue the option leading to the greatest expected value. The idea that agents can compare options based on expected value has motivated a search for neural representations of value predictions in the brain, an endeavor that has been enormously fruitful (for some caveats, see O'Doherty 2014). Value signals have been found in a range of brain regions, including the amygdala, orbitofrontal cortex (OFC), ventromedial prefrontal cortex (vmPFC), and ventral and dorsal striata, as well as in a number of other brain areas such as the parietal, premotor, and dorsal frontal areas.

Reinforcement Learning

Evidence for value signals in the brain raises the question of how such signals could be learned or acquired in the first place. The seminal work of Schultz and colleagues (1997) has provided insight into a potential mechanism for signal learning; they found that the phasic activity of dopamine neurons encodes a prediction error, which signals the difference between expected and actual rewards. Referred to as a reward prediction error (RPE), phasic dopamine activity has been shown to resemble, both in signature and function, a signal used by computational reinforcement learning (RL) algorithms to support learning (Montague et al. 1996, Sutton 1988). This type of learning signal allows an agent to improve its prediction of what to expect from the environment by continually adjusting those predictions toward what actually occurred. The fact that dopamine neurons send dense projections to the striatum and elsewhere has given rise to proposals that RPE signals carried by phasic dopamine facilitate neural plasticity associated with the acquisition of value predictions in these target areas.

Model-Free and Model-Based Reinforcement Learning

A flurry of interest followed the realization that abstract learning theories from computer science could be applied to better understand the brain at a computational level within a RL framework (Doya 1999). In particular, Daw and colleagues (2005) proposed that the distinction between habitual and goal-directed control could be accounted for in terms of two distinct types of RL mechanisms.

When learning is mediated via RPE signals, value is ascribed only by integrating across past reinforcement. Predictive value acquired via this mechanism does not include the agent's motivation at the time of reinforcement, nor does it track the identity of the reinforcer itself. Thus, a controller that learns via RPE signals would be expected to behave in a manner that is insensitive to immediate changes in outcome values, similar to the devaluation insensitivity associated with habits. In essence, this model-free learning strategy (so called because it does not depend on a model of the environment) gives rise to value representation that resembles stimulus-based association.

To account for goal-directed control, Daw and colleagues (2005) proposed that the agent encodes an internal model of the decision problem consisting of the relevant states and actions and, critically, the transition structure among them. This map of the decision process supports flexible online value computation by considering the current expected value of

outcomes and integrating into these expected values the knowledge of how to procure them. Critically, value can be flexibly constructed at each decision point as part of an online planning procedure, making the agent immediately sensitive to changes in outcome values. This type of cognitive model–driven RL process is known, perhaps somewhat confusingly (because the terms were originally coined in the computer science literature), as model-based RL (Kuvayev & Sutton 1996).

NEUROCOMPUTATIONAL SUBSTRATES

Formal RL algorithms depend on well-defined learning signals and representations. Therefore, by asking how these are implemented in the brain, we can move toward a better understanding of the brain's computational composition. In the following sections, we outline some of the key representations and signals associated with various forms of RL and discuss their neural correlates. Figure 1 illustrates the main brain regions and functions discussed in these sections.

The Cognitive Model: Multiple Maps, Multiple Regions

A model-based agent depends on a cognitive map of the task space encoding the environment's relevant features and the relationships among them (Tolman 1948). Electrophysiological recordings from place cells in the hippocampus have provided the most well-characterized evidence for the encoding of a cognitive map, especially in the spatial domain (e.g., O'Keefe & Dostrovsky 1971). Activity in these cells can represent the animal's trajectory during a spatial decision-making task, consistent with the theory that place cell representations play a role in model-based planning (Pfeiffer & Foster 2013) and that place cells are recruited in correspondence with future spatial locations the animal is considering (Johnson & Redish 2007). Others have suggested that the hippocampus might play a more general role in encoding a cognitive map, possibly in the encoding of relationships between stimuli and outcomes, identity and category membership information about objects (Eichenbaum et al. 1999), or even maps of social hierarchy in humans (Tavares et al. 2015).

Although evidence suggests that the hippocampus encodes information relevant to a cognitive map, the hippocampus does not always seem to be necessary for goal-directed choices in simple action–outcome learning tasks (Corbit & Balleine 2000). Wilson et al. (2014) used computational modeling to account for various behavioral effects of orbitofrontal lesions in the extant literature and to suggest that the OFC is involved in signaling the current location of the animal in an abstract task space, especially when that state is not immediately observable (i.e., when task states must be inferred or maintained). Neuroimaging studies have revealed evidence that outcome identity is represented in the OFC in response to stimuli predictive of those outcomes (Howard et al. 2015). This representation may be a mechanism through which the expected value of a particular stimulus or state could be computed. Although this possibility is still a matter of debate, the bulk of the evidence suggests that the OFC seems to be less involved in encoding information about actions than it is in encoding information about stimuli and outcomes (for a review, see Rangel & Hare 2010). Ultimately, the OFC's role in state encoding and in

outcome associations may ultimately service computations associated with the expected value based on stimulus–stimulus associations.

However, goal-directed action selection demands some form of action representation as well as a representation of the state transitions afforded by performing actions. Evidence has indicated that regions of the posterior parietal cortex, such as the lateral intraparietal sulcus, play an important role in perceptual decision making, a critical aspect of state identification (e.g., Shadlen & Newsome 2001). Notably, neurons in the posterior parietal cortex have been implicated in the encoding of information about stimulus category membership, which could be important for establishing current and future potential states (Freedman & Assad 2006). Indeed, work by Doll et al. (2015) has shown that the category of a prospective stimulus appears to engage these regions of the brain. Critically, neurons in the posterior parietal cortex are implicated in the encoding of associations between arbitrary stimuli; these associations indicate the implementation of specific actions (Dorris & Glimcher 2004). A region of the inferior parietal lobule has also been found to play an important role in the encoding of information pertinent to the distribution of outcomes associated with an action, as well as information about the relative probability of obtaining an outcome contingent on performing a particular action compared to not performing that action (Liljeholm et al. 2011, 2013). Together, these findings suggest a role for the posterior parietal cortex in encoding a cognitive map or, more specifically, in encoding the transitions between states contingent on specific actions.

The presence of cognitive maps in the brain raises the question of how such maps are acquired in the first place. One possible mechanism is a state prediction error (SPE), which signals the discrepancy between an expected state transition and the transition that actually did occur. This SPE can then be used to adjust state transition expectations. In essence, SPEs are similar to RPEs but are used not to learn about reward expectation but to learn state expectations. Gläscher et al. (2010) used fMRI while participants learned a two-step Markov decision problem to find evidence for SPEs in the posterior parietal cortex and dorsolateral prefrontal cortex. These SPE signals were present in both a latent learning task phase, during which participants were guided through the task in the absence of reward, and an active phase, during which reward, and therefore RPEs, were also present. SPEs in the posterior parietal cortex and dorsolateral prefrontal cortex are therefore candidates for the signal underpinning learning of a cognitive model involving actions.

The presence of multiple candidate areas engaged in encoding some form of a cognitive map raises the question of which representations are necessary or sufficient for model-based learning and control. The nature of the cognitive map representation that is used may depend to a great extent on the type of decision problem. Perhaps, a task that has an ostensibly spatial component will necessarily recruit a spatial cognitive map in the hippocampus, whereas decision problems that involve selection among possible motor actions will depend to a greater extent on action codes in the posterior parietal cortex. However, precisely how these various maps might be leveraged by the brain in support of model-based learning and control remains to be determined.

Outcome Valuation During Decision Making

To choose among actions in a model-based manner, an agent needs to determine the value of different available outcomes. Electrophysiological studies in both rodents and monkeys have revealed neuronal activity in the amygdala and OFC related to conditioned stimuli associated with appetitive unconditioned stimuli, such as a sweet taste or juice reward (Schoenbaum et al. 1998), and aversive unconditioned stimuli, such as an aversive flavor, air puff, or eyelid shock (Applegate et al. 1982, Pascoe & Kapp 1985, Paton et al. 2006, Salzman & Fusi 2010, Salzman et al. 2007, Schoenbaum et al. 1998). Furthermore, human imaging studies have revealed responses in the amygdala, ventral striatum, and OFC 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; Tobler et al. 2006).

During Pavlovian conditioning, many of these brain areas are involved in triggering Pavlovian conditioned responses. The central nucleus of the amygdala projects to lateral hypothalamic and brainstem nuclei involved in implementing conditioned autonomic reflexes (LeDoux et al. 1988). The ventral striatum sends projections via the globus pallidus to motor nuclei in the brainstem, such as the pedunculopontine nucleus (Groenewegen & Berendse 1994, Winn et al. 1997). This projection pattern is compatible with a possible role for the ventral striatum in triggering conditioned skeletomotor reflexes, such as approach and avoidance behavior, as well as consummatory responses. As we discuss in the section Action Valuation and Planning, the output of this network of brain areas is also taken into consideration by a separate network of brain areas when organisms have to choose among different actions in order to gain a desired outcome. First, we explore in greater detail the representations and signals carried by some of these areas.

Value signals have been found in both the OFC and the vmPFC. Electrophysiological recordings in area 13 of the central OFC of nonhuman primates revealed that neurons in this area encode the value of differing amounts of juice on offer (Padoa-Schioppa & Assad 2006). The activity of some of these neurons correlated with the subjective value of each of the two outcomes on offer, whereas other neurons correlated with the subjective value of the outcome that was ultimately chosen. Rodent studies have found similar results, with value signals associated with expected delivery of an outcome being present in the rodent OFC (McDannald et al. 2011, Schoenbaum et al. 1998). Other neurophysiological studies of monkeys have reported neuronal responses correlating with the value of prospective outcomes throughout the OFC and in other brain regions, including the lateral prefrontal and anterior cingulate cortices (Lee et al. 2007, Seo et al. 2007, Smith et al. 2010, Wallis & Miller, 2003). Interestingly, neurons in the lateral prefrontal cortex have been found to respond in a manner consistent with the outcome value associated with novel stimuli whose value must be inferred from the outcome of the previous trial, suggesting that these value representations are sensitive to higher-order task structure (Pan et al. 2014). The vmPFC of humans seems to encode similar representations. Activity in the vmPFC was found to correlate with trial-by-trial variations in the amount participants were willing to pay (WTP) for offered goods (Plassmann et al. 2007). A follow-up experiment comparing value representations for foods, which participants would pay to obtain or avoid, revealed vmPFC

activity proportional to the value of goods with positive values and decreasing activity scaling with negative values (Plassmann et al. 2010).

Organisms are forced to choose not only among rewards of varying probability and magnitude but also among rewards that differ in type. Organisms may cope with this issue by representing and comparing outcome values in a common currency. Indeed, activity in overlapping regions of the vmPFC correlated with the subjective value of three distinct categories of goods in a WTP task: food items, nonfood consumer items, and money (Chib et al. 2009). Levy & Glimcher (2012) found evidence for a common currency in the vmPFC by giving participants explicit choices between different types of goods, specifically money versus food, and by demonstrating that activation levels scaled according to the common currency value for both types of good. Although these findings are consistent with the notion of a common currency, they could also be the result of averaging nonoverlapping value representations across individual subjects if there is sufficient spatial variance in these representations among individuals. Using a paradigm similar to that of Chib et al. (2009), McNamee et al. (2013) probed for distributed voxel patterns encoding outcome value and category by training multivariate pattern classifiers on each type of good. A circumscribed region of the vmPFC above the orbital surface was found to exhibit a general value code whereby a classifier trained on the value of one class of goods (e.g., foods) could successfully decode the value of goods from a different category (e.g., consumer goods). In addition to general value codes, value codes specific to particular categories of good were also found along the medial orbital surface, a finding that is consistent with the idea that these regions represent value in a preliminary category-specific form that is then converted into a common currency in more dorsal parts of the vmPFC. Interestingly, no region was found to uniquely encode the distributed value of monetary items, which were only found to be represented in the vmPFC, perhaps because money is a generalized reinforcer that can be exchanged for many different types of goods.

Taken together, these findings support the existence of a common currency in the vmPFC in which the value of various outcomes are proportionally scaled in accordance with subjective value irrespective of the category from which they are drawn. In the following section, we consider how other information relevant to model-based computations is encoded.

Outcome Valuation After a Decision Has Been Made

In addition to evaluating outcomes while forming a decision, an organism also has to evaluate an outcome once it has been received. Extensive evidence implicates the vmPFC and adjacent parts of the OFC in the response to experienced outcomes, including monetary rewards (Knutson et al. 2001, O'Doherty et al. 2001, Smith et al. 2010); taste, odor, and flavor (de Araujo et al. 2003a,b, Rolls et al. 2003); attractive faces (O'Doherty et al. 2003a); and the aesthetic value of abstract art (Kirk et al. 2009). These outcome representations are also strongly influenced by changes in underlying motivational states. The vmPFC and OFC show decreasing responses to food, odor, or even water outcomes as motivational states change from hungry or thirsty to satiated, paralleling changes in the subjective pleasantness of the stimulus (de Araujo et al. 2003a,b, O'Doherty et al. 2000, Rolls et al. 2003, Small et al. 2001). Not only are such representations modulated as a function of changes in internal

motivational state, but value-related activity in this region is also influenced by cognitive factors, such as the provision of price information or even the mere use of semantic labels (de Araujo et al. 2005, Plassmann et al. 2008). Thus, the online computation of outcome values in the vmPFC and OFC is highly flexible and influenced by a variety of internal and external factors.

Action Valuation and Planning

Once an organism has determined the value of different outcomes, it must often determine the value of available actions based on how likely they are to lead to a desired outcome. To calculate these so-called model-based action values, a decision-making agent must be armed with a cognitive map that will enable the retrieval of probability distributions over the future states or outcomes that can be attained. The model-free computation of action value, i.e., computation without any consideration of state transitions or of which outcome might be achieved, is discussed in the section Neurobiological Substrates of Model-Free Action Selection.

One strategy for calculating model-based action values involves iteration over states, actions, and state transitions. Given that model-based action values depend on arithmetic computations accounting for quantity and probability, brain systems traditionally associated with working memory, such as the lateral prefrontal cortex (Miller & Cohen 2001), as well as parts of the parietal cortex implicated in numerical cognition (Platt & Glimcher 1999), are likely to be involved. It therefore seems reasonable to hypothesize that regions of the frontal and parietal cortices play a fundamental role in the computation of model-based action values. In a result that is at least partly consistent with this possibility, Simon & Daw (2011) reported increasing activity in the dorsolateral prefrontal and anterior cingulate cortices as a function of the depth of model-based planning during a spatial navigation task. In addition, areas of the posterior parietal cortex are also important in action planning. Distinct neuronal populations seem to be specialized for planning particular actions (such as saccades versus reaching movements), and these neurons appear to be specifically involved in encoding action trajectories and representing the target state of the action trajectories in both monkeys (Andersen et al. 1997, Cohen & Andersen 2002, MacKay 1992) and humans (Desmurget et al. 1999).

In rodents, several studies have produced evidence for a distinct network of brain areas supporting goal-directed behavior. Evidence from these studies indicates that the prelimbic cortex, as well as the dorsomedial striatum in the basal ganglia, to which the prelimbic cortex projects, are involved in the acquisition of goal-directed responses. Studies in rodents show that lesions to these areas impair action–outcome learning, rendering the rodent's behavior permanently stimulus-driven (Baker & Ragozzino 2014, Balleine & Dickinson 1998, Ragozzino et al. 2002, Yin et al. 2005). Although the prelimbic cortex is involved in the initial acquisition of goal-directed learning, this region does not appear to be essential for the expression of goal-directed actions after acquisition (Ostlund & Balleine 2005). In contrast, the dorsomedial striatum appears to be necessary for both acquisition and expression of goal-directed behavior (Yin et al. 2005).

Some researchers have argued that the rodent prelimbic cortex and dorsomedial striatum correspond to the primate vmPFC and caudate nucleus, respectively (Balleine & O'Doherty 2009). Indeed, in addition to representing the value of the different outcomes on offer (as discussed in the previous section), activity in the vmPFC also tracks instrumental contingencies, i.e., the causal relationship between an action and an outcome, sensitivity to which has also been shown to be associated with goal-directed control in rodent studies (Liljeholm et al. 2011, Matsumoto et al. 2003). Contingency manipulations have also implicated the caudate nucleus in goal-directed behavior in nonhuman primates (Hikosaka et al. 1989) and humans (Liljeholm et al. 2011). Furthermore, activity in the vmPFC has been found to track the current incentive value of an instrumental action such that, following devaluation, activity decreases for an action associated with a devalued outcome relative to an action associated with a still-valued outcome (de Wit et al. 2009, Valentin et al. 2007). Interestingly, the strength of the connection between the vmPFC and dorsomedial striatum as measured with diffusion tensor imaging has been shown to correlate with the degree of goal-directed behavioral expression across individuals (de Wit et al. 2012).

Once action values have been computed, they can be compared at decision points. Although several studies have reported evidence for prechoice action values, few studies have determined whether or not such action-value representations are computed in a model-based or model-free manner. Studies in rodents and monkeys report action-value signals in the dorsal striatum, as well as in areas of the dorsal cortex, including the parietal and supplementary motor cortices (Kolb et al. 1994, Lau & Glimcher 2008, Platt & Glimcher 1999, Samejima et al. 2005, Sohn & Lee 2007, Whitlock et al. 2012, Wilber et al. 2014). Human fMRI studies report evidence that putative action-value signals are present in areas of the dorsal cortex, including the supplementary motor, lateral parietal, and dorsolateral cortices (Hare et al. 2011, Morris et al. 2014, Wunderlich et al. 2009).

Little is known about how organisms integrate the range of variables that appear to influence action selection. One candidate region for the site of this integration is the dorsomedial prefrontal cortex. In monkeys, Hosokawa and colleagues (2013) found that some neurons in the anterior cingulate cortex are involved in encoding an integrated value signal that summed over expected costs and benefits for an action. Hunt et al. (2014) also implicated a region of the dorsomedial prefrontal cortex in encoding integrated action values. Together, these preliminary findings support the possibility that action valuation involves an interaction between multiple brain systems and that goal-value representations in the vmPFC are ultimately integrated with action information in dorsal cortical regions to compute an overall action value.

Neurobiological Substrates of Model-Free Action Selection

The canonical learning signal implicated in model-free value learning is the RPE, which is thought to be encoded by the phasic activity of midbrain dopamine neurons (Schultz et al. 1997). Evidence indicates that reward-related prediction errors also play a role in learning in humans. Numerous fMRI studies have reported correlations between RPE signals from RL models and activity in the striatum and midbrain nuclei known to contain dopaminergic

neurons during Pavlovian and instrumental learning paradigms (D'Ardenne et al. 2008, O'Doherty 2004, O'Doherty et al. 2003b,Pauli et al. 2015, Wittmann et al. 2005).

Other evidence suggests that the dorsal striatum is critical for learning the stimulus–response associations underlying habitual behavior. In rodents, lesions of the posterior dorsolateral striatum have been found to render behavior permanently goal-directed such that, after overtraining, these animals fail to express habits (Yin et al. 2004, 2006). Tricomi et al. (2009) demonstrated a link between increasing activity in the human posterior striatum as a function of training and the emergence of habitual control as assessed with a reinforcer devaluation test. Wunderlich et al. (2012) reported that activity in this area correlated with the value of overtrained actions (which might be expected to favor habitual control) compared to actions whose values had been acquired more recently. Others have reported putative model-free value signals in the posterior putamen (Horga et al. 2015).

The phasic activity of dopamine neurons is causally related to learning of instrumental actions via dopamine-modulated plasticity in target areas of these neurons, such as the dorsolateral striatum (Faure et al. 2005, Schoenbaum et al. 2013, Steinberg & Janak 2013). Human fMRI studies of motor sequence learning have reported an increase in activity in the posterior dorsolateral striatum as sequences become overlearned. For instance, participants who successfully learn to perform instrumental actions for reward show significantly stronger prediction error signals in the dorsal striatum than those who fail to learn instrumental actions (Schönberg et al. 2007), and the administration of drugs that modulate dopamine function, such as L-3,4-dihydroxyphenylalanine (L-DOPA) or dopaminergic antagonists, influences the strength of learning of instrumental associations accordingly (Frank et al. 2004). Other studies focusing on both model-based and model-free value signals have also found evidence for model-free signals in the posterior putamen (Doll et al. 2015, Lee et al. 2014). However, model-free signals have also been reported across a number of cortical areas (Lee et al. 2014). Moreover, differences in the strength of the connectivity between the right posterolateral striatum and the premotor cortex across individuals is associated with differences in the degree to which individuals show evidence of habitual behavior in a task in which goal-directed and habitual responses are placed in conflict (de Wit et al. 2012).

Other Decision Variables: Effort and Uncertainty

One variable that is likely to play an important role during decision making is the amount of effort, whether cognitive or physical, involved in performing a particular action. Clearly, all else being equal, it is better to exert as little effort as possible, but occasions may arise in which effortful actions yield disproportionately greater rewards. Although effort studies are scarce, there is evidence that the effort associated with performing an action is represented in parts of the dorsomedial prefrontal cortex alongside other areas such as the insular cortex (Prévost et al. 2010). Additional studies in rodents suggest that the anterior cingulate cortex plays a critical role in effortful behavior (Hillman & Bilkey 2012, Walton et al. 2009).

Two forms of uncertainty, expected and estimation uncertainty, may also be relevant factors at the time of decision. The most pertinent form of expected uncertainty for decision making is risk, or the inherent stochasticity of the environment that remains even when the

contingencies are fully known. Expected uncertainty regarding different options is useful information to access at the point of decision making because risk preference might vary over time depending on motivational and other contextual factors. Studies have revealed activity correlating with expected uncertainty in a number of cortical and subcortical brain regions, including the insular cortex, inferior frontal gyrus, and dorsal striatum (Critchley et al. 2001, Huettel et al. 2006, Paulus et al. 2003, Yanike & Ferrera 2014).

In contrast to risk, estimation uncertainty corresponds to uncertainty in the estimate of the reward distribution associated with a particular action or state. For example, the first time an action is sampled in a particular context, estimation uncertainty is high; it will decrease as that action is repeated and the precision of the reward distribution's estimate increases. Estimation uncertainty can also be leveraged to balance the trade-off between exploration and exploitation by allowing the agent to target actions that are relatively undersampled. Neural representations of estimation uncertainty have been reported in the anterior cingulate cortex (Payzan-LeNestour et al. 2013), and uncertainty signals (which may or may not correspond to estimation uncertainty) associated with exploration have also been reported in the frontopolar cortex (Badre et al. 2012, Daw et al. 2006, Yoshida & Ishii 2006).

Model-Free and Model-Based Pavlovian Learning

In this section, we turn our attention to the computations that underpin acquisition and expression of Pavlovian conditioned responses. As described in the section Neurobiological Substrates of Model-Free Action Selection, model-free RL has been proposed as a mechanism to underpin learning in at least appetitive Pavlovian conditioning. However, similar to the predictions in the instrumental domain, a model-free RL account of Pavlovian conditioning would be expected to produce conditioned responses that are devaluation insensitive. Nevertheless, many conditioned Pavlovian responses are strongly devaluation sensitive (Dayan & Berridge 2014). This discrepancy has led to suggestions that model-based learning mechanisms might also apply in the case of Pavlovian conditioning (Dayan & Berridge 2014, Prévost et al. 2013).

We might expect such a system to depend on a cognitive model that maps the relationship between different stimuli, that is, a model that encodes stimulus–stimulus association likelihoods. One might expect the mechanism for model-based Pavlovian conditioning to be similar to that involved in model-based instrumental control, with the exception that there is no need for the model to represent action contingencies. Sensory preconditioning represents one piece of behavioral evidence in favor of the existence of a model-based Pavlovian learning mechanism that depends on the formation of stimulus–stimulus associations. In sensory preconditioning, two cues are repeatedly paired together in the absence of reward. Following this, one of the cues is paired with reward. Rescorla (1980) found that, under these conditions, the cue that had not been paired with reward also spontaneously elicited appetitive conditioned responses (Rescorla 1980).

This result raises the question of which brain areas are involved in encoding stimulus–stimulus associations. The hippocampus and the OFC, which we have examined in the context of their role in encoding a cognitive map, are strong candidates. Representations in these two brain regions are perhaps not action dependent but do encode relationships

between stimuli, as would be needed by a model-based Pavlovian mechanism. Indeed, consistent with this proposal, both the hippocampus and OFC are implicated in sensory preconditioning (Jones et al. 2012, Holland & Bouton 1999, Wimmer & Shohamy 2012). Researchers have also found that the amygdala encodes information about context, stimulus identity, and reward expectation (Salzman & Fusi 2010). Moreover, Prévost et al. (2013) used a Pavlovian reversal learning paradigm to provide evidence for expected value signals in the human amygdala that were better captured by a model-based algorithm than by a number of model-free learning alternatives.

Two distinct forms of Pavlovian appetitive conditioning, sign tracking and goal tracking, can be distinguished in rodents (Boakes 1977, Hearst & Jenkins 1974, Jenkins & Moore 1973). Signtracking animals orient to the cue that predicts the subsequent reward, whereas goal-tracking animals orient to the location where the outcome is delivered. A recent behavioral study has revealed a correlation between the extent to which animals manifest sign-tracking behavior and the extent to which these animals show evidence of devaluation insensitivity in their behavior, suggesting that sign tracking may be a model-free conditioned response (Nasser et al. 2015). Consistent with dopamine's involvement in model-free Pavlovian conditioning, RPE signals in the nucleus accumbens core have been associated with sign tracking. Animals selectively bred to be predominantly sign trackers show phasic dopamine release in the nucleus accumbens, whereas animals bred to be predominantly goal trackers do not show clear phasic dopaminergic activity during learning (Flagel et al. 2007). Furthermore, a recent study has found evidence to suggest that phasic dopaminergic activity associated with a conditioned stimulus may in fact be devaluation insensitive, as would be predicted by a model-free algorithm. Specifically, rats were conditioned to associate a cue with an aversive salt outcome. Following induction of a salt appetite, dopamine neurons showed increased phasic activity following the receipt of the (now-valued) salt outcome, consistent with model-based control. However, consistent with a model-free RL mechanism, phasic responses to the cue predicting salt did not show any such increase until after the animal had a chance to be exposed to the outcome, suggesting that dopamine activity in response to the cue was not immediately updated to reflect the current value of the associated outcome (Cone et al. 2016). These findings suggest that in Pavlovian conditioning, dopaminergic prediction errors may be involved in model-free but not model-based learning.

INTERACTION AMONG BEHAVIORAL CONTROL SYSTEMS

Having considered evidence regarding the existence of multiple control systems in the brain and reviewed ideas and emerging evidence about the possible neural computations underpinning each of these systems, we briefly consider in the following sections how these systems interact. There is evidence to suggest that stimulus-driven, goal-directed, and noninstrumental systems may sometimes interact in an adaptive manner whereby each system exerts complementary influences on behavior in a manner beneficial for the agent. Alternatively, in some instances these systems can interact in a maladaptive manner, leading to pernicious behavioral outcomes.

Interactions Between Goals and Habits

Habitual and goal-directed control systems may interact to provide a strategy that is both flexible and cognitively efficient by supporting hierarchical decomposition of the task at hand. Building on theoretical work demonstrating the computational benefits of encapsulating behavioral invariance in the form of a selectable option (Sutton et al. 1999), studies have begun to probe whether the brain leverages its varied control systems to implement a similar hierarchical decomposition (Botvinick 2012, Botvinick et al. 2009). Evidence from human fMRI studies shows that higher levels of abstraction progressively engage more anterior regions of frontal cortex, suggesting a hierarchical organization of abstraction along a rostral–caudal axis (Badre & D'Esposito 2007, Donoso et al. 2014, Koechlin et al. 2003). Other studies have reported signals consistent with hierarchical event structuring (Schapiro et al. 2013) and prediction errors (Diuk et al. 2013, Ribas-Fernandes et al. 2011). Although the most common depiction of hierarchical control positions the stimulus-driven system as subservient to the goal-directed system (Dezfouli & Balleine 2013), other work suggests that the goal-directed system can also be deployed in the service of a habitually selected goal (Cushman & Morris 2015).

The brain's multiple control systems may also facilitate learning. Situations in which control is assigned to the goal-directed system in the early stages of behavioral acquisition may be examples of adaptive interactions between systems. Once the problem space has been sufficiently sampled, behavioral control transitions to the habitual system, thereby freeing up cognitive resources that would otherwise be allocated to the goal-directed system. The complementary nature of the interactions between these systems is such that, even though the goal-directed system is in the driving seat during early learning, the habitual system is given the opportunity to learn a model-free policy because it is exposed to the relevant stimulus associations.

However, there is a downside to this training interaction. Once behavior is under the control of the habitual system, it may guide the agent toward an unfavorable course of action under circumstances in which environmental contingencies have changed or the agent's goals have changed. Alternatively, errors in goal-directed representations may inculcate inappropriate biases into the stimulus-driven system's learned values (Doll et al. 2011). Numerous examples of maladaptive interactions exist in the realm of psychiatric disease. For instance, habits for abuse of a drug may persist even if the goal of the individual is to stop taking the drug (Everitt & Robbins 2016). Overeating or compulsive behaviors may also be examples of the habitual system exerting inappropriate and ultimately detrimental control over behavior (Voon et al. 2015). The capacity to effectively manage conflicting policy suggestions by the goal-directed and habitual systems likely varies across individuals and may even relate to underlying differences in the neural circuitry, perhaps indicative of differing levels of vulnerability to the emergence of compulsive behavior (de Wit et al. 2012).

Interactions with Pavlovian Predictions

The Pavlovian system can also interact with systems involved in instrumental behavior, a class of interactions referred to as Pavlovian-to-instrumental transfer (PIT) (Lovibond 1983).

PIT effects are typically manifested as increased instrumental response vigor in the presence of a reward predicting a Pavlovian conditioned stimulus (Estes 1943). One can make a distinction between general and specific PIT. General PIT refers to circumstances in which a Pavlovian cue motivates increased instrumental responding irrespective of the outcome associated with the Pavlovian cue. Conversely, outcome-specific PIT effects modulate responding when both the Pavlovian cue and instrumental action are associated with the same outcome (Corbit & Balleine 2005, Holland & Gallagher 2003, Rescorla & Solomon 1967).

In a normative relationship between incentives and instrumental response, the provision of higher incentives should result in increased effort and response accuracy, thereby enabling more effective action implementation. However, Pavlovian effects on instrumental responding can also promote maladaptive behavior in circumstances in which PIT effects continue to exert an energizing effect on instrumental actions associated with a devalued outcome (Holland 2004, Watson et al. 2014; although see Allman et al. 2010). This suggests that PIT effects selectively involve the habitual system. Thus, Pavlovian cues may intervene in the interplay between goals and habits by actively biasing behavioral control toward the habitual system.

Furthermore, under certain circumstances, increased incentives can paradoxically result in less-efficacious instrumental performance, an effect known as choking that has been linked to dopaminergic regions of the midbrain (Chib et al. 2014, Mobbs et al. 2009, Zedelius et al. 2011). For example, Ariely et al. (2009) offered participants in rural India the prospect of winning large monetary amounts relative to their average monthly salaries. Compared to a group offered smaller incentive amounts, the performance of the high-incentive group was much impaired, suggesting the counterintuitive effect of reduced performance in a situation in which the motivation to succeed is high. Numerous theories have been proposed to account for choking effects, reflecting various possible forms of interactions between different control systems. One theory is that choking effects reflect a maladaptive return of behavioral control to the goal-directed system in the face of large potential incentives in a situation in which the habitual system is better placed to reliably execute a skilled behavior. Although some results support this hypothesis (Lee & Grafton 2015), others support an alternative account whereby Pavlovian effects elicited by cues could engage Pavlovian skeletomotor behaviors, such as appetitive approach or aversive withdrawal, that interfere with the performance of the habitual skilled motor behavior (Chib et al. 2012, 2014). More than one of these ideas could hold true, as behavioral choking effects may have multiple causes arising from maladaptive interactions between these systems.

Arbitration Between Behavioral Control Mechanisms

The presence of distinct control systems burdens the brain with the problem of how to apportion control among them. An influential hypothesis is that there exists an arbitrator that determines the influence each system has over behavior based on a number of criteria (Daw et al. 2005). One important factor in this hypothesis is the relative accuracy of the systems' predictions concerning which action should be selected; all else being equal, behavior should be controlled by the system with the most accurate prediction (Daw et al. 2005).

Using the computational distinction between model-based and model-free RL, Lee et al. (2014) found evidence for the existence of an arbitration processes in the ventrolateral prefrontal cortex and frontopolar cortex that assigns behavioral control as a function of system reliability. Connectivity between the arbitration areas and the regions of the brain encoding habitual but not goal-directed action values was also found to be modulated as a function of the arbitration process. Consistent with a default model-free strategy, it is better to delegate control to the more-efficient stimulus-driven system; however, when the arbitration system detects that a goal-directed policy is warranted, then it may achieve this through active inhibition of the habitual system, leaving the model-based system free to control behavior. In addition to predictive accuracy, other relevant variables include the amount of cognitive effort required (FitzGerald et al. 2014) and the potential benefits that can be accrued by implementing a model-based strategy (Pezzulo et al. 2013, Shenhav et al. 2013).

Much less is known about how arbitration occurs between Pavlovian and instrumental systems. Changes in cognitive strategies or appraisal implemented via the prefrontal cortex can influence the likelihood of both aversive and appetitive Pavlovian conditioned responses, perhaps via downregulation of the amygdala and ventral striatum (Delgado et al. 2008a,b; Staudinger et al. 2009). This type of top-down process could be viewed as a form of arbitration, in which Pavlovian control policies are downweighted in situations in which goal-directed control is deemed to be more beneficial. However, the nature of the computations mediating this putative arbitration process is not well understood. Clearly, given that Pavlovian behaviors are often advantageous in time-critical situations when the animal's survival may be at stake, it would be reasonable for at least certain types of Pavlovian predictions to have immediate access to behavior without having to wait for the arbitration process to mediate. Therefore, it seems plausible to expect that, perhaps as with the habitual system, arbitration operates only to inhibit Pavlovian behavior when it is deemed to be inappropriate or irrelevant. One might also predict that any such arbitration process would happen at a relatively slower timescale relative to the more rapid response time available to the Pavlovian system. Therefore, traces of initial Pavlovian control might become manifest in behavior even in situations in which the arbitration system subsequently implements an inhibition of the Pavlovian system.

Neural Systems for Learning and Inference in a Social Context

Thus far, we have considered the involvement of multiple systems in controlling reward-related behavior but have given scant attention to the type of behavioral context in which these systems are engaged. A particularly challenging problem faced by humans and many other animals is the need to learn from and ultimately behave adaptively to conspecifics. Succinctly put, the problem is working out how to conduct oneself in social situations. A full consideration of this issue is beyond the scope of this review. However, we can briefly consider the question of whether value-based action selection in social contexts depends on similar or distinct control systems and neural circuitry as those involved in value-based action selection in nonsocial contexts.

One of the simplest ways to extend the framework we have discussed to the social domain is to apply this framework to the mechanisms underlying observational learning, which allow an agent to learn about the value of stimuli or actions not through direct experience but instead through observing the behavior of another agent. Several studies have revealed the engagement of brain regions including the ventral and dorsal striata and the vmPFC in observational learning (Burke et al. 2010, Cooper et al. 2012). For example, Cooper et al. (2012) found evidence for prediction error signals in the striatum when participants were learning about the value of actions through observing another agent. These preliminary findings suggest that, at least for some forms of observational learning, the brain relies on similar neural mechanisms and circuitry for learning through observation as it does when learning through direct experience. There is also evidence to suggest that, during a number of social situations in which it is necessary to learn from the actions being taken by others, the brain may rely on similar circuitry and updating signals as those known to be involved in model-based RL (Abe & Lee 2011, Liljeholm et al. 2012, Seo et al. 2009).

However, in some social situations, the brain may engage additional circuitry that has been implicated in mentalizing or theory of mind (Frith & Frith 2003, 2006). For instance, Hampton et al. (2008) found that when participants engage in a competitive game against a dynamic opponent, activity in the posterior superior temporal sulcus and dorsomedial prefrontal cortex is related to the updating of a higher-order inference about the strategic intentions of that opponent. Relatedly, Behrens et al. (2008) examined a situation in which it was useful for participants to learn about the reliability of a confederate's recommendations about what actions to take because the confederate's interests sometimes lay in deceiving the subject. Neural activity corresponding to an update signal for such an estimate was found in the anterior medial prefrontal cortex, as well as in a region of the temporoparietal junction. Similarly, Boorman et al. (2013a,b) found evidence for updating signals related to learning about another individual's expertise on a financial investment task in the temporoparietal junction and dorsomedial frontal cortex. Suzuki et al. (2015) found evidence for the representation of beliefs about the likely future actions of a group of individuals in the posterior superior temporal sulcus and, moreover, found that this activity was specifically engaged when performing in a social as compared to a nonsocial context.

Taken together, these findings suggest that, although learning and making decisions in a social context often depends on similar brain circuitry as that used when learning in nonsocial contexts, additional distinct circuitry is deployed to facilitate socially relevant tasks, such as inferring the internal mental states of others, when knowledge about relevant features of another agent is necessary.

CONCLUSIONS AND FUTURE DIRECTIONS

Although much remains to be explored, the past few decades have brought considerable advances in our understanding of the neural and computational mechanisms underlying learning, reward, and decision making. Merging formal work in computational intelligence and empirical research in cognitive neuroscience has allowed considerable headway not only in understanding the algorithms embodied by the brain but also in illuminating how the brain navigates the trade-offs between different strategies for controlling reward-related behavior.

Long-standing theoretical arguments as to whether behavior is habitual or goal-directed have been assuaged by demonstrations that the brain has maintained multiple strategies for behavioral control, each offering advantages and disadvantages that may be leveraged across a range of potential circumstances.

As a result of these advances, new unresolved issues have emerged. In this article, we have reviewed evidence from both animal and human studies indicating that a goal-directed (model-based) system guides behavior in some circumstances but that other situations favor a habitual (model-free) strategy. Factors such as task familiarity, task complexity, and reward contingencies may influence the trade-off between these two systems; however, work remains to be done regarding other variables that might influence how various strategies are deployed. Factors such as incentives (the benefits of favoring one strategy over another), cognitive capacities (the brain's awareness of its own limitations), and social context may play a role in system deployment. Whether Pavlovian drives factor into the arbitration scheme used to determine behavioral control also remains unknown.

Furthermore, we understand little regarding the mechanisms through which system arbitration is instantiated. We have presented evidence suggesting that the brain adopts a computationally efficient model-free strategy by default but that this can be interrupted by a more flexible goal-directed strategy if needed. However, this evidence raises the question of what the model-based system is doing when it is not favored for control: Is the model-based system passively working in the background, waiting to be called back into activity, or has it moved offline to conserve resources? If the latter, how is it brought back online in a sensible way? We must also ask what the model-free system is doing when the model-based system takes control. There is evidence to suggest that the model-based system can shape the model-free system's value representations, but we know very little about this relationship. Does the model-free system passively learn about choices and experiences governed by the model-based system, or can the model-based system tutor the model-free system more directly and, if so, how might this be operationalized?

The bulk of our discussion has focused on behavioral control with respect to what can be labeled as exploitive action selection: identifying and moving toward the most rewarding options in the environment. However, this is only one half of what is commonly referred to as the explore/exploit trade-off. Almost nothing is known about the role played by the brain's varied control systems with respect to exploration. Given the exploitive advantages that come with having multiple control strategies, some of which we have outlined in this review, at one's disposal, are similar benefits offered to the domain of exploration? Does the brain take advantage of the computational efficiencies offered by the model-free system to direct exploration, or does the novelty and complexity inherent to exploration demand a model-based strategy? Perhaps multiple strategies are deployed in a collaborative fashion to tackle the many facets of exploration in an efficient way. Issues pertinent to the brain's engagement with exploratory decision making are ripe for both theoretical and experimental research.

Finally, we briefly touched upon the role played by the brain's control systems in a social context. However, the nature of these additional learning and inference signals and how they

interact with other control systems is not yet fully understood. Value signals in the vmPFC and anterior cingulate cortex do reflect knowledge of strategic information and the information needed to modify the value signals to reflect this knowledge appears to arrive via inputs from the mentalizing network (Hampton et al. 2008, Suzuki et al. 2015). Whether these mentalizing-related computations can be considered a fourth system for guiding behavior or, instead, a module that provides input into the model-based system is an open question. Moreover, how the brain decides when or whether the mentalizing system should be engaged in a particular situation is currently unknown, although it is tempting to speculate that an arbitration process may play a role.

This, of course, is only a small sample of many questions the field of decision neuroscience is poised to tackle. Although pursuit of these issues will deepen our basic understanding of the brain's functional architecture, of equal importance will be our ability to apply these concepts toward our understanding of cognitive impairments and mental illness (Huys et al. 2016, Maia & Frank 2011, Montague et al. 2012). Despite many advances and huge incentives, and perhaps in testament to the complexity of the problem, reliable and effective treatments are scarce. By building on a functional understanding of the brain's learning and control strategies, their points of interaction, and the mechanisms by which they manifest, novel treatments (whether behavioral, chemical, or mechanistic) may be able to help millions of people lead more fulfilling lives.

ACKNOWLEDGMENTS

This work was supported by a National Institutes of Health (NIH) Conte Center grant for research on the neurobiology of social decision making (P50MH094258-01A1), NIH grant number DA033077-01 (supported by OppNet, NIH's Basic Behavioral and Social Science Opportunity Network), and National Science Foundation grant number 1207573 to J.O.D.

LITERATURE CITED

Abe H, Lee D. 2011 Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. Neuron 70(4):731–41 [PubMed: 21609828]

Allman MJ, DeLeon IG, Cataldo MF, Holland PC, Johnson AW. 2010 Learning processes affecting human decision making: an assessment of reinforcer-selective Pavlovian-to-instrumental transfer following reinforcer devaluation. J. Exp. Psychol. Anim. Behav. Process 36(3):402–8 [PubMed: 20658871]

Andersen RA, Snyder LH, Bradley DC, Xing J. 1997 Multimodal representation of space in the posterior parietal cortex and its use in planning movements. Annu. Rev. Neurosci 20:303–30 [PubMed: 9056716]

Applegate CD, Frysinger RC, Kapp BS, Gallagher M. 1982 Multiple unit activity recorded from amygdala central nucleus during Pavlovian heart rate conditioning in rabbit. Brain Res 238(2):457–62 [PubMed: 7093668]

Ariely D, Gneezy U, Loewenstein G, Mazar N. 2009 Large stakes and big mistakes. Rev. Econ. Stud 76(2):451–69

Badre D, D'Esposito M. 2007 Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. J. Cogn. Neurosci 19(12):2082–99 [PubMed: 17892391]

Badre D, Doll BB, Long NM, Frank MJ. 2012 Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73(3):595–607 [PubMed: 22325209]

Baker PM, Ragozzino ME. 2014 Contralateral disconnection of the rat prelimbic cortex and dorsomedial striatum impairs cue-guided behavioral switching. Learn. Mem 21(8):368–79 [PubMed: 25028395]

Balleine BW, Daw ND, O'Doherty JP. 2009 Multiple forms of value learning and the function of dopamine. In Glimcher et al. 2013, pp. 367–85

Balleine BW, Dickinson A. 1991 Instrumental performance following reinforcer devaluation depends upon incentive learning. Q. J. Exp. Psychol. Sect. B 43(3):279–96

Balleine BW, Dickinson A. 1998 Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37(4–5):407–19 [PubMed: 9704982]

Balleine BW, O'Doherty JP. 2009 Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35(1):48–69

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. 2008 Associative learning of social value. Nature 456(7219):245–49 [PubMed: 19005555]

Boakes RA. 1977 Performance on learning to associate a stimulus with positive reinforcement In Operant-Pavlovian Interactions, ed. Davis H, Burwitz HMB, pp. 67–97. London: Wiley

Boorman ED, O'Doherty JP, Adolphs R, Rangel A. 2013a The behavioral and neural mechanisms underlying the tracking of expertise. Neuron 80(6):1558–71 [PubMed: 24360551]

Boorman ED, Rushworth MF, Behrens TE. 2013b Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. J. Neurosci 33(6):2242–53 [PubMed: 23392656]

Botvinick MM. 2012 Hierarchical RL and decision making. Curr. Opin. Neurobiol 22(6):956–62 [PubMed: 22695048]

Botvinick MM, Niv Y, Barto AC. 2009 Hierarchically organized behavior and its neural foundations: a RL perspective. Cognition 113(3):262–80 [PubMed: 18926527]

Burke CJ, Tobler PN, Baddeley M, Schultz W. 2010 Neural mechanisms of observational learning. PNAS 107(32):14431–36 [PubMed: 20660717]

Camerer C, Loewenstein G, Prelec D. 2005 Neuroeconomics: How neuroscience can inform economics. J. Econ. Lit 43:9–64

Chib VS, De Martino B, Shimojo S, O'Doherty JP. 2012 Neural mechanisms underlying paradoxical performance for monetary incentives are driven by loss aversion. Neuron 74(3):582–94 [PubMed: 22578508]

Chib VS, Rangel A, Shimojo S, O'Doherty JP. 2009 Evidence for a common representation of decision values for dissimilar goods in human VmPFC. J. Neurosci 29(39):12315–20 [PubMed: 19793990]

Chib VS, Shimojo S, O'Doherty JP. 2014 The effects of incentive framing on performance decrements for large monetary outcomes: behavioral and neural mechanisms. J. Neurosci 34(45):14833–44 [PubMed: 25378151]

Cohen YE, Andersen RA. 2002 A common reference frame for movement plans in the posterior parietal cortex. Nat. Rev. Neurosci 3(7):553–62 [PubMed: 12094211]

Cone JJ, Fortin SM, McHenry JA, Stuber GD, McCutcheon JE, Roitman MF. 2016 Physiological state gates acquisition and expression of mesolimbic reward prediction signals. PNAS 113(7):1943–48 [PubMed: 26831116]

Cooper JC, Dunne S, Furey T, O'Doherty JP. 2012 Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. J. Cogn. Neurosci 24(1):106–18 [PubMed: 21812568]

Corbit LH, Balleine BW. 2000 The role of the hippocampus in instrumental conditioning. J. Neurosci 20(11):4233–39 [PubMed: 10818159]

Corbit LH, Balleine BW. 2005 Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian-instrumental transfer. J. Neurosci 25(4):962–70 [PubMed: 15673677]

Critchley HD, Mathias CJ, Dolan RJ. 2001 Neural activity in the human brain relating to uncertainty and arousal during anticipation. Neuron 29(2):537–45 [PubMed: 11239442]

Cushman F, Morris A. 2015 Habitual control of goal selection in humans. PNAS 112(45):13817–22 [PubMed: 26460050]

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD. 2008 BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319(5867):1264–67 [PubMed: 18309087]

Davey GCL. 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 ND, Niv Y, Dayan P. 2005 Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci 8(12):1704–11 [PubMed: 16286932]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. 2006 Cortical substrates for exploratory decisions in humans. Nature 441(7095):876–79 [PubMed: 16778890]

Dayan P, Berridge KC. 2014 Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. Cogn. Affect. Behav. Neurosci 14(2):473–92 [PubMed: 24647659]

De Araujo IET, Kringelbach ML, Rolls ET, McGlone F. 2003a Human cortical responses to water in the mouth, and the effects of thirst. J. Neurophysiol 90(3):1865–76 [PubMed: 12773496]

De Araujo IET, Rolls ET, Kringelbach ML, McGlone F, Phillips N. 2003b Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain. Eur. J. Neurosci 18(7):2059–68 [PubMed: 14622239]

De Araujo IET, Rolls ET, Velazco MI, Margot C, Cayeux I. 2005 Cognitive modulation of olfactory processing. Neuron 46(4):671–79 [PubMed: 15944134]

de Wit S, Corlett PR, Aitken MR, Dickinson A, Fletcher PC. 2009 Differential engagement of the VmPFC by goal-directed and habitual behavior toward food pictures in humans. J. Neurosci 29(36):11330–38 [PubMed: 19741139]

de Wit S, Watson P, Harsay HA, Cohen MX, van de Vijver I, Ridderinkhof KR. 2012 Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. J. Neurosci 32(35):12066–75 [PubMed: 22933790]

Delgado MR, Li J, Schiller D, Phelps EA. 2008a The role of the striatum in aversive learning and aversive prediction errors. Philos. Trans. R. Soc. Lond. B Biol. Sci 363(1511):3787–800 [PubMed: 18829426]

Delgado MR, Nearing KI, Ledoux JE, Phelps EA. 2008b Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. Neuron 59(5):829–38 [PubMed: 18786365]

Desmurget M, Epstein CM, Turner RS, Prablanc C, Alexander GE, Grafton ST. 1999 Role of the posterior parietal cortex in updating reaching movements to a visual target. Nat. Neurosci 2(6): 563–67 [PubMed: 10448222]

Dezfouli A, Balleine BW. 2013 Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLOS Comput. Biol 9(12):e1003364 [PubMed: 24339762]

Dickinson A 1985 Actions and habits: the development of behavioural autonomy. Philos. Trans. R. Soc. Lond. B Biol. Sci 308(1135):67–78

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

Dickinson A, Balleine B, Watt A, Gonzalez F, Boakes RA 1995 Motivational control after extended instrumental training. Anim. Learn. Behav 23(2):197–206

Dickinson A, Nicholas DJ, Adams CD. 1983 The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. Q. J. Exp. Psychol. Sect. B 35(1):35–51

Diuk C, Tsai K, Wallis J, Botvinick M, Niv Y. 2013 Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. J. Neurosci 33(13):5797–805 [PubMed: 23536092]

Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND. 2015 Model-based choices involve prospective neural activity. Nat. Neurosci 18(5):767–72 [PubMed: 25799041]

Doll BB, Hutchison KE, Frank MJ. 2011 Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J. Neurosci 31(16):6188–98 [PubMed: 21508242]

Donoso M, Collins AGE, Koechlin E. 2014 Foundations of human reasoning in the prefrontal cortex. Science 344(6191):1481–86 [PubMed: 24876345]

Dorris MC, Glimcher PW. 2004 Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron 44(2):365–78 [PubMed: 15473973]

Doya K 1999 What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw 12(7–8):961–74 [PubMed: 12662639]

Eichenbaum H, Dudchenko P, Wood E, Shapiro M, Tanila H. 1999 The hippocampus, memory, and place cells: Is it spatial memory or a memory space? Neuron 23(2):209–26 [PubMed: 10399928]

Estes WK. 1943 Discriminative conditioning. I. A discriminative property of conditioned anticipation. J. Exp. Psychol 32(2):150–55

Everitt BJ, Robbins TW. 2016 Drug addiction: updating actions to habits to compulsions ten years on. Annu. Rev. Psychol 67(1):23–50 [PubMed: 26253543]

Faure A, Haberland U, Condé F, Massioui NE 2005 Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. J. Neurosci 25(11):2771–80 [PubMed: 15772337]

FitzGerald THB, Dolan RJ, Friston KJ. 2014 Model averaging, optimal inference, and habit formation. Front. Hum. Neurosci 8:457 [PubMed: 25018724]

Flagel SB, Watson SJ, Robinson TE, Akil H. 2007 Individual differences in the propensity to approach signals versus goals promote different adaptations in the dopamine system of rats. Psychopharmacol. Berl 191(3):599–607

Frank MJ, Seeberger LC, O'Reilly RC. 2004 By carrot or by stick: cognitive RL in parkinsonism. Science 306(5703):1940–43 [PubMed: 15528409]

Freedman DJ, Assad JA. 2006 Experience-dependent representation of visual categories in parietal cortex. Nature 443(7107):85–88 [PubMed: 16936716]

Frith CD, Frith U. 2006 The neural basis of mentalizing. Neuron 50(4):531–34 [PubMed: 16701204]

Frith U, Frith CD. 2003 Development and neurophysiology of mentalizing. Philos. Trans. R. Soc. Lond. B Biol. Sci 358(1431):459–73 [PubMed: 12689373]

Gigerenzer G, Gaissmaier W. 2011 Heuristic decision making. Annu. Rev. Psychol 62(1):451–82 [PubMed: 21126183]

Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010 States versus rewards: dissociable neural prediction error signals underlying model-based and model-free RL. Neuron 66(4):585–95 [PubMed: 20510862]

Glimcher PW, Camerer CF, Fehr E, Poldrack RA, eds. 2013 Neuroeconomics: Decision Making and the Brain. London: Academic

Gottfried JA, O'Doherty J, Dolan RJ. 2002 Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. J. Neurosci 22(24):10829–37 [PubMed: 12486176]

Gottfried JA, O'Doherty J, Dolan RJ. 2003 Encoding predictive reward value in human amygdala and OFC. Science 301(5636):1104–7 [PubMed: 12934011]

Groenewegen HJ, Berendse HW. 1994 Anatomical relationships between the prefrontal cortex and the basal ganglia in the rat In Motor and Cognitive Functions of the Prefrontal Cortex, ed. Thierry AM, Glowinski J, Goldman-Rakic PS, Christen Y, pp. 51–77. Berlin/Heidelberg: Springer

Hampton AN, Bossaerts P, O'Doherty JP. 2008 Neural correlates of mentalizing-related computations during strategic interactions in humans. PNAS 105(18):6741–46 [PubMed: 18427116]

Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A. 2011 Transformation of stimulus value signals into motor commands during simple choice. PNAS 108(44):18120–25 [PubMed: 22006321]

Hearst E, Jenkins HM. 1974 Sign-Tracking: The Stimulus-Reinforcer Relation and Directed Action. Madison, WI: Psychon. Soc.

Hikosaka O, Sakamoto M, Usui S. 1989 Functional properties of monkey caudate neurons. I. Activities related to saccadic eye movements. J. Neurophysiol 61(4):780–98 [PubMed: 2723720]

Hillman KL, Bilkey DK. 2012 Neural encoding of competitive effort in the anterior cingulate cortex. Nat. Neurosci 15(9):1290–97 [PubMed: 22885851]

Holland PC. 2004 Relations between Pavlovian-instrumental transfer and reinforcer devaluation. J. Exp. Psychol. Anim. Behav. Process 30(2):104–17 [PubMed: 15078120]

Holland PC, Bouton ME. 1999 Hippocampus and context in classical conditioning. Curr. Opin. Neurobiol 9(2):195–202 [PubMed: 10322181]

Holland PC, Gallagher M. 2003 Double dissociation of the effects of lesions of basolateral and central amygdala on conditioned stimulus-potentiated feeding and Pavlovian-instrumental transfer. Eur. J. Neurosci 17(8):1680–94 [PubMed: 12752386]

Horga G, Maia TV, Marsh R, Hao X, Xu D, et al. 2015 Changes in corticostriatal connectivity during RL in humans. Hum. Brain Mapp 36(2):793–803 [PubMed: 25393839]

Hosokawa T, Kennerley SW, Sloan J, Wallis JD. 2013 Single-neuron mechanisms underlying cost-benefit analysis in frontal cortex. J. Neurosci 33(44):17385–97 [PubMed: 24174671]

Howard JD, Gottfried JA, Tobler PN, Kahnt T. 2015 Identity-specific coding of future rewards in the human orbitofrontal cortex. PNAS 112(16):5195–200 [PubMed: 25848032]

Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML. 2006 Neural signatures of economic preferences for risk and ambiguity. Neuron 49(5):765–75 [PubMed: 16504951]

Hunt LT, Dolan RJ, Behrens TEJ. 2014 Hierarchical competitions subserving multi-attribute choice. Nat. Neurosci 17(11):1613–22 [PubMed: 25306549]

Huys QJM, Maia TV, Frank MJ. 2016 Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci 19(3):404–13 [PubMed: 26906507]

Jenkins HM, Moore BR. 1973 The form of the auto-shaped response with food or water reinforcers. J. Exp. Anal. Behav 20(2):163–81 [PubMed: 4752087]

Johnson A, Redish AD. 2007 Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci 27(45):12176–89 [PubMed: 17989284]

Jones JL, Esber GR, McDannald MA, Gruber AJ, Hernandez A, et al. 2012 OFC supports behavior and learning using inferred but not cached values. Science 338(6109):953–56 [PubMed: 23162000]

Kirk U, Skov M, Hulme O, Christensen MS, Zeki S. 2009 Modulation of aesthetic value by semantic context: an fMRI study. NeuroImage 44(3):1125–32 [PubMed: 19010423]

Knutson B, Fong GW, Adams CM, Varner JL, Hommer D. 2001 Dissociation of reward anticipation and outcome with event-related fMRI. Neuroreport 12(17):3683–87 [PubMed: 11726774]

Koechlin E, Ody C, Kouneiher F. 2003 The architecture of cognitive control in the human prefrontal cortex. Science 302(5648):1181–85 [PubMed: 14615530]

Kolb B, Buhrmann K, McDonald R, Sutherland RJ. 1994 Dissociation of the medial prefrontal, posterior parietal, and posterior temporal cortex for spatial navigation and recognition memory in the rat. Cereb. Cortex 4(6):664–80 [PubMed: 7703691]

Konorski J, Miller S. 1937 On two types of conditioned reflex. J. Gen. Psychol 16(1):264–72

Kuvayev L, Sutton R. 1996 Model-based RL with an approximate, learned model. Proc. Yale Worksh. Adapt. Learn. Syst., 9th, June 10–12, New Haven, CT, pp. 101–5. New Haven, CT: Dunham Lab., Yale Univ.

Lau B, Glimcher PW 2008 Value representations in the primate striatum during matching behavior. Neuron 58(3):451–63 [PubMed: 18466754]

LeDoux JE, Iwata J, Cicchetti P, Reis DJ. 1988 Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. J. Neurosci 8(7):2517–29 [PubMed: 2854842]

Lee D, Rushworth MFS, Walton ME, Watanabe M, Sakagami M. 2007 Functional specialization of the primate frontal cortex during decision making. J. Neurosci 27(31):8170–73 [PubMed: 17670961]

Lee SW, Shimojo S, O'Doherty JP. 2014 Neural computations underlying arbitration between model-based and model-free learning. Neuron 81(3):687–99 [PubMed: 24507199]

Lee TG, Grafton ST. 2015 Out of control: Diminished prefrontal activity coincides with impaired motor performance due to choking under pressure. NeuroImage 105:145–55 [PubMed: 25449744]

Levy DJ, Glimcher PW. 2012 The root of all value: a neural common currency for choice. Curr. Opin. Neurobiol 22(6):1027–38 [PubMed: 22766486]

Liljeholm M, Molloy CJ, O'Doherty JP. 2012 Dissociable brain systems mediate vicarious learning of stimulus-response and action-outcome contingencies. J. Neurosci 32(29):9878–86 [PubMed: 22815503]

Liljeholm M, Tricomi E, O'Doherty JP, Balleine BW. 2011 Neural correlates of instrumental contingency learning: differential effects of action-reward conjunction and disjunction. J. Neurosci 31(7):2474–80 [PubMed: 21325514]

Liljeholm M, Wang S, Zhang J, O'Doherty JP 2013 Neural correlates of the divergence of instrumental probability distributions. J. Neurosci 33(30):12519–27 [PubMed: 23884955]

Lovibond PF. 1983 Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. J. Exp. Psychol. Anim. Behav. Process 9(3):225–47 [PubMed: 6153052]

MacKay WA. 1992 Properties of reach-related neuronal activity in cortical area 7A. J. Neurophysiol 67(5):1335–45 [PubMed: 1597716]

Maia TV, Frank MJ. 2011 From RL models to psychiatric and neurological disorders. Nat. Neurosci 14(2):154–62 [PubMed: 21270784]

Matsumoto K, Suzuki W, Tanaka K. 2003 Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science 301(5630):229–32 [PubMed: 12855813]

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. 2011 Ventral striatum and OFC are both required for model-based, but not model-free, RL. J. Neurosci 31(7):2700–5 [PubMed: 21325538]

McNamee D, Rangel A, O'Doherty JP. 2013 Category-dependent and category-independent goal-value codes in human vmPFC. Nat. Neurosci 16(4):479–85 [PubMed: 23416449]

Miller EK, Cohen JD. 2001 An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci 24(1):167–202 [PubMed: 11283309]

Mobbs D, Hassabis D, Seymour B, Marchant JL, Weiskopf N, et al. 2009 Choking on the money: Reward-based performance decrements are associated with midbrain activity. Psychol. Sci 20(8): 955–62 [PubMed: 19594859]

Montague PR, Dayan P, Sejnowski TJ 1996 A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci 16(5):1936–47 [PubMed: 8774460]

Montague PR, Dolan RJ, Friston KJ, Dayan P. 2012 Computational psychiatry. Trends Cogn. Sci 16(1):72–80 [PubMed: 22177032]

Morris RW, Dezfouli A, Griffiths KR, Balleine BW 2014 Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. Nat. Commun 5:4390 [PubMed: 25055179]

Nasser HM, Chen Y-W, Fiscella K, Calu DJ. 2015 Individual variability in behavioral flexibility predicts sign-tracking tendency. Front. Behav. Neurosci 9:289 [PubMed: 26578917]

O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C. 2001 Abstract reward and punishment representations in the human OFC. Nat. Neurosci 4(1):95–102 [PubMed: 11135651]

O'Doherty J, Rolls ET, Francis S, Bowtell R, McGlone F, et al. 2000 Sensory-specific satiety-related olfactory activation of the human OFC. Neuroreport 11(4):893–97 [PubMed: 10757540]

O'Doherty J, Winston J, Critchley H, Perrett D, Burt DM, Dolan RJ. 2003a Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. Neuropsychologia 41(2):147–55 [PubMed: 12459213]

O'Doherty JP. 2004 Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr. Opin. Neurobiol 14(6):769–76 [PubMed: 15582382]

O'Doherty JP. 2014 The problem with value. Neurosci. Biobehav. Rev 43:259–68 [PubMed: 24726573]

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. 2003b Temporal difference models and reward-related learning in the human brain. Neuron 38(2):329–37 [PubMed: 12718865]

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ. 2002 Neural responses during anticipation of a primary taste reward. Neuron 33(5):815–26 [PubMed: 11879657]

O'Keefe J, Dostrovsky J. 1971 The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res 34(1):171–75 [PubMed: 5124915]

Ostlund SB, Balleine BW. 2005 Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. J. Neurosci 25(34):7763–70 [PubMed: 16120777]

Padoa-Schioppa C, Assad JA. 2006 Neurons in the OFC encode economic value. Nature 441(7090): 223–26 [PubMed: 16633341]

Pan X, Fan H, Sawa K, Tsuda I, Tsukada M, Sakagami M. 2014 Reward inference by primate prefrontal and striatal neurons. J. Neurosci 34(4):1380–96 [PubMed: 24453328]

Pascoe JP, Kapp BS. 1985 Electrophysiological characteristics of amygdaloid central nucleus neurons during Pavlovian fear conditioning in the rabbit. Behav. Brain Res 16(2–3):117–33 [PubMed: 4041212]

Paton JJ, Belova MA, Morrison SE, Salzman CD. 2006 The primate amygdala represents the positive and negative value of visual stimuli during learning. Nature 439(7078):865–70 [PubMed: 16482160]

Pauli WM, Larsen T, Collette S, Tyszka JM, Seymour B, O'Doherty JP. 2015 Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. J. Neurosci 35(42):14220–33 [PubMed: 26490862]

Paulus MP, Rogalsky C, Simmons A, Feinstein JS, Stein MB. 2003 Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. NeuroImage 19(4):1439–48 [PubMed: 12948701]

Pavlov I 1927 Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. London: Oxford Univ. Press

Payzan-LeNestour E, Dunne S, Bossaerts P, O'Doherty JP. 2013 The neural representation of unexpected uncertainty during value-based decision making. Neuron 79(1):191–201 [PubMed: 23849203]

Pezzulo G, Rigoli F, Chersi F. 2013 The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. Front. Psychol 4:92 [PubMed: 23459512]

Pfeiffer BE, Foster DJ. 2013 Hippocampal place-cell sequences depict future paths to remembered goals. Nature 497(7447):74–79 [PubMed: 23594744]

Plassmann H, O'Doherty J, Rangel A. 2007 OFC encodes willingness to pay in everyday economic transactions. J. Neurosci 27(37):9984–88 [PubMed: 17855612]

Plassmann H, O'Doherty JP, Rangel A. 2010 Appetitive and aversive goal values are encoded in the medial OFC at the time of decision making. J. Neurosci 30(32):10799–808 [PubMed: 20702709]

Plassmann H, O'Doherty J, Shiv B, Rangel A. 2008 Marketing actions can modulate neural representations of experienced pleasantness. PNAS 105(3):1050–54 [PubMed: 18195362]

Platt ML, Glimcher PW. 1999 Neural correlates of decision variables in parietal cortex. Nature 400(6741):233–38 [PubMed: 10421364]

Prévost C, McNamee D, Jessup RK Bossaerts P, O'Doherty JP. 2013 Evidence for model-based computations in the human amygdala during Pavlovian conditioning. PLOS Comput. Biol 9(2):e1002918 [PubMed: 23436990]

Prévost C, Pessiglione M, Météreau E, Cléry-Melin M-L, Dreher J-C. 2010 Separate valuation subsystems for delay and effort decision costs. J. Neurosci 30(42):14080–90 [PubMed: 20962229]

Ragozzino ME, Ragozzino KE, Mizumori SJ, Kesner RP. 2002 Role of the dorsomedial striatum in behavioral flexibility for response and visual cue discrimination learning. Behav. Neurosci 116(1):105–15 [PubMed: 11898801]

Rangel A, Camerer C, Montague PR. 2008 A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci 9(7):545–56 [PubMed: 18545266]

Rangel A, Hare T. 2010 Neural computations associated with goal-directed choice. Curr. Opin. Neurobiol 20(2):262–70 [PubMed: 20338744]

Rescorla RA. 1980 Simultaneous and successive associations in sensory preconditioning. J. Exp. Psychol. Anim. Behav. Process 6(3):207–16 [PubMed: 6153051]

Rescorla RA, Solomon RL. 1967 Two-process learning theory: relationships between Pavlovian conditioning and instrumental learning. Psychol. Rev 74(3):151–82 [PubMed: 5342881]

Rescorla RA, Wagner AR. 1972 A theory of Pavlovian conditioning: variations in the effectiveness of re-inforcement and nonreinforcement In Classical Conditioning II: Current Research and Theory, ed. Black AH, Prokasy WF, pp. 64–99. New York: Appleton-Century-Crofts

Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, et al. 2011 A neural signature of hierarchical RL. Neuron 71(2):370–79 [PubMed: 21791294]

Rolls ET, Kringelbach ML, De Araujo IET. 2003 Different representations of pleasant and unpleasant odours in the human brain. Eur. J. Neurosci 18(3):695–703 [PubMed: 12911766]

Salzman CD, Fusi S. 2010 Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. Annu. Rev. Neurosci 33:173–202 [PubMed: 20331363]

Salzman CD, Paton JJ, Belova MA, Morrison SE. 2007 Flexible neural representations of value in the primate brain. Ann. N. Y. Acad. Sci 1121(1):336–54 [PubMed: 17872400]

Samejima K, Ueda Y, Doya K, Kimura M. 2005 Representation of action-specific reward values in the striatum. Science 310(5752):1337–40 [PubMed: 16311337]

Schapiro AC, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM. 2013 Neural representations of events arise from temporal community structure. Nat. Neurosci 16(4):486–92 [PubMed: 23416451]

Schoenbaum G, Chiba AA, Gallagher M. 1998 OFC and basolateral amygdala encode expected outcomes during learning. Nat. Neurosci 1(2):155–59 [PubMed: 10195132]

Schoenbaum G, Esber GR, Iordanova MD. 2013 Dopamine signals mimic reward prediction errors. Nat. Neurosci 16(7):777–79 [PubMed: 23799468]

Schönberg T, Daw ND, Joel D, O'Doherty JP. 2007 Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J. Neurosci 27(47):12860–67 [PubMed: 18032658]

Schultz W, Dayan P, Montague PR 1997 A neural substrate of prediction and reward. Science 275(5306):1593–99 [PubMed: 9054347]

Seo H, Barraclough DJ, Lee D. 2007 Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb. Cortex 17(Suppl. 1):110–17

Seo H, Barraclough DJ, Lee D. 2009 Lateral intraparietal cortex and RL during a mixed-strategy game. J. Neurosci 29(22):7278–89 [PubMed: 19494150]

Shadlen MN, Newsome WT. 2001 Neural basis of a perceptual decision in the parietal cortex (Area LIP) of the rhesus monkey. J. Neurophysiol 86(4):1916–36 [PubMed: 11600651]

Shenhav A, Botvinick MM, Cohen JD. 2013 The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79(2):217–40 [PubMed: 23889930]

Simon DA, Daw ND. 2011 Neural correlates of forward planning in a spatial decision task in humans. J. Neurosci 31(14):5526–39 [PubMed: 21471389]

Small DM, Zatorre RJ, Dagher A, Evans AC, Jones-Gotman M. 2001 Changes in brain activity related to eating chocolate. Brain 124(9):1720–33 [PubMed: 11522575]

Smith DV, Hayden BY, Truong T-K, Song AW, Platt ML, Huettel SA. 2010 Distinct value signals in anterior and posterior VmPFC. J. Neurosci 30(7):2490–95 [PubMed: 20164333]

Sohn J-W, Lee D. 2007 Order-dependent modulation of directional signals in the supplementary and pre-supplementary motor areas. J. Neurosci 27(50):13655–66 [PubMed: 18077677]

Stalnaker TA, Franz TM, Singh T, Schoenbaum G. 2007 Basolateral amygdala lesions abolish orbitofrontal-dependent reversal impairments. Neuron 54(1):51–58 [PubMed: 17408577]

Staudinger MR, Erk S, Abler B, Walter H. 2009 Cognitive reappraisal modulates expected value and prediction error encoding in the ventral striatum. NeuroImage 47(2):713–21 [PubMed: 19442745]

Steinberg EE, Janak PH. 2013 Establishing causality for dopamine in neural function and behavior with optogenetics. Brain Res 1511:46–64 [PubMed: 23031636]

Strait CE, Blanchard TC, Hayden BY. 2014 Reward value comparison via mutual inhibition in VmPFC. Neuron 82(6):1357–66 [PubMed: 24881835]

Sutton RS. 1988 Learning to predict by the methods of temporal differences. Mach. Learn 3(1):9–44

Sutton RS 1990 RL architectures for animats. Proc. Int. Conf. Simul. Adapt. Behav., 1st, From Animals to Animats, Cambridge, MA, pp. 288–96. Cambridge, MA: MIT Press

Sutton RS, Precup D, Singh S. 1999 Between MDPs and semi-MDPs: a framework for temporal abstraction in RL. Artif. Intell 112:181–211

Suzuki S, Adachi R, Dunne S, Bossaerts P, O'Doherty JP. 2015 Neural mechanisms underlying human consensus decision-making. Neuron 86(2):591–602 [PubMed: 25864634]

Tavares RM, Mendelsohn A, Grossman Y, Williams CH, Shapiro M, et al. 2015 A map for social navigation in the human brain. Neuron 87(1):231–43 [PubMed: 26139376]

Thibodeau GA, Patton KT. 1992 Structure & Function of the Body. St. Louis, MO: Mosby Year Book. 9th ed.

Thorndike EL. 1898 Animal intelligence: an experimental study of the associative processes in animals. Psychol. Rev. Monogr. Suppl 2(4):1–109

Tobler PN, O'Doherty JP, Dolan RJ, Schultz W. 2006 Human neural learning depends on reward prediction errors in the blocking paradigm. J. Neurophysiol 95(1):301–10 [PubMed: 16192329]

Tolman EC. 1948 Cognitive maps in rats and men. Psychol. Rev 55(4):189–208 [PubMed: 18870876]

Tricomi E, Balleine BW, O'Doherty JP. 2009 A specific role for posterior dorsolateral striatum in human habit learning. Eur. J. Neurosci 29(11):2225–32 [PubMed: 19490086]

Tully T, Quinn WG. 1985 Classical conditioning and retention in normal and mutant *Drosophila melanogaster*. J. Comp. Physiol 157(2):263–77 [PubMed: 3939242]

Valentin VV, Dickinson A, O'Doherty JP. 2007 Determining the neural substrates of goal-directed learning in the human brain. J. Neurosci 27(15):4019–26 [PubMed: 17428979]

Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, et al. 2015 Disorders of compulsivity: a common bias towards learning habits. Mol. Psychiatry 20(3):345–52 [PubMed: 24840709]

Wallis JD, Miller EK. 2003 Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. Eur. J. Neurosci 18(7):2069–81 [PubMed: 14622240]

Walters ET, Carew TJ, Kandel ER. 1981 Associative learning in *Aplysia*: evidence for conditioned fear in an invertebrate. Science 211(4481):504–6 [PubMed: 7192881]

Walton ME, Groves J, Jennings KA, Croxson PL, Sharp T, et al. 2009 Comparing the role of the anterior cingulate cortex and 6-hydroxydopamine nucleus accumbens lesions on operant effort-based decision making. Eur. J. Neurosci 29(8):1678–91 [PubMed: 19385990]

Watson P, Wiers RW, Hommel B, de Wit S. 2014 Working for food you don't desire. Cues interfere with goal-directed food-seeking. Appetite 79:139–48 [PubMed: 24743030]

Whitlock JR, Pfuhl G, Dagslott N, Moser M-B, Moser EI. 2012 Functional split between parietal and entorhinal cortices in the rat. Neuron 73(4):789–802 [PubMed: 22365551]

Wilber AA, Clark BJ, Forster TC, Tatsuno M, McNaughton BL. 2014 Interaction of egocentric and world-centered reference frames in the rat posterior parietal cortex. J. Neurosci 34(16):5431–46 [PubMed: 24741034]

Wilson RC, Takahashi YK, Schoenbaum G, Niv Y. 2014 OFC as a cognitive map of task space. Neuron 81(2):267–79 [PubMed: 24462094]

Wimmer GE, Shohamy D. 2012 Preference by association: how memory mechanisms in the hippocampus bias decisions. Science 338(6104):270–73 [PubMed: 23066083]

Winn P, Brown VJ, Inglis WL. 1997 On the relationships between the striatum and the pedunculopontine tegmental nucleus. Crit. Rev. Neurobiol 11(4):241–61 [PubMed: 9336713]

Wittmann BC, Schott BH, Guderian S, Frey JU, Heinze H-J, Düzel E 2005 Reward-related fMRI activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. Neuron 45(3):459–67 [PubMed: 15694331]

Wunderlich K, Rangel A, O'Doherty JP. 2009 Neural computations underlying action-based decision making in the human brain. PNAS 106(40):17199–204 [PubMed: 19805082]

Wunderlich K, Dayan P, Dolan RJ. 2012 Mapping value based planning and extensively trained choice in the human brain. Nat. Neurosci 15(5):786–91 [PubMed: 22406551]

Yanike M, Ferrera VP. 2014 Representation of outcome risk and action in the anterior caudate nucleus. J. Neurosci 34(9):3279–90 [PubMed: 24573287]

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):181–89 [PubMed: 14750976]

Yin HH, Knowlton BJ, Balleine BW. 2005 Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. Eur. J. Neurosci 22(2):505–12 [PubMed: 16045503]

Yin HH, Knowlton BJ, Balleine BW. 2006 Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. Behav. Brain Res 166(2):189–96 [PubMed: 16153716]

Yoshida W, Ishii S. 2006 Resolution of uncertainty in prefrontal cortex. Neuron 50(5):781–89 [PubMed: 16731515]

Zedelius CM, Veling H, Aarts H. 2011 Boosting or choking—how conscious and unconscious reward processing modulate the active maintenance of goal-relevant information. Conscious. Cogn 20(2):355–62 [PubMed: 20510630]

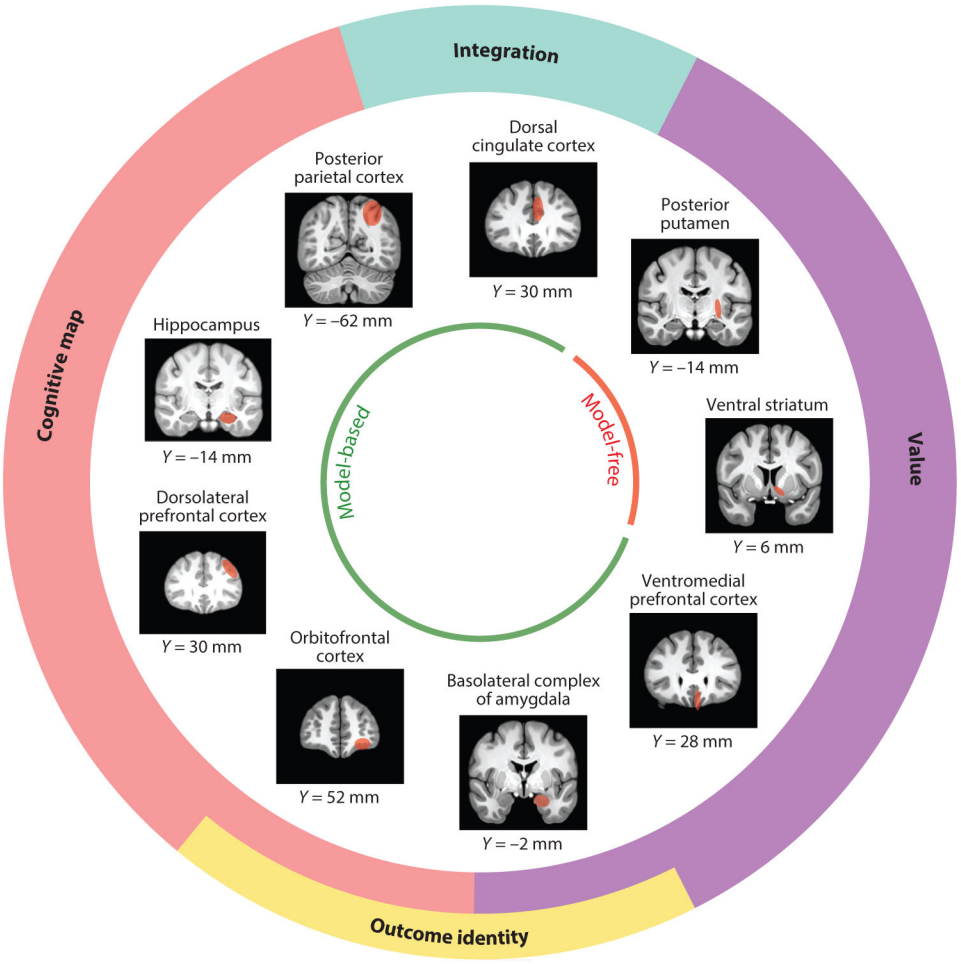

Figure 1.

Schematic mapping specific neuroanatomical loci to the implementation of different functions underlying model-based and model-free control. Model-based control depends on a cognitive map of state space and integration of different aspects of a decision, such as effort and estimation uncertainty, as well as the value and the identity of goals or outcomes. Model-free control depends on learning about the value of responses in the current state, based on the history of past reinforcement. The inner circle identifies regions involved in model-based and model-free control, and the outer circle identifies specific subfunctions implemented by particular brain regions, based on the evidence to date as discussed in this review. The objective of this figure is to orient the reader to the location of the relevant brain regions rather than to provide a categorical description of the functions of each region or an exhaustive list of the brain regions involved in reward-related behavior. The neuronal substrates of prediction errors and the loci of arbitration mechanisms are omitted from this figure for simplicity. *Y* coordinates of coronal brain slices represent their distance from the commissures along the posterior (negative values) to anterior (positive values) axis.