Learning, Reward and Decision-Making

John P. O'Doherty[1], Jeffrey Cockburn[1*], Wolfgang M. Pauli[1*]

[1]Division of Humanities and Social Sciences and Computation and Neural Systems Program, California Institute of Technology, Pasadena, California 91125

corresponding author: John P. O'Doherty, MC 228-77, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125

Email address: J.P. O'Doherty <jdoherty@caltech.edu>; Jeffrey Cockburn <jcockbur@caltech.edu>; Wolfgang M. Pauli <pauli@caltech.edu>
*These authors contributed equally

Contents

Keywords

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

Abstract

We summarize findings supporting the existence of multiple behavioral strategies for controlling reward-related behavior, including a goal-directed or model-based system, a 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, showing how these can lead to either adaptive or maladaptive behavioral outcomes. We then review evidence that an additional system guides inference over 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.

Introduction

A fundamental challenge faced by all organisms, including humans, concerns 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. A variety of strategies to solve this problem have emerged over the course of evolution. There is accumulating evidence to suggest that these distinct strategies co-exist in the human brain. In this review, we will outline evidence for the existence of these multiple systems of behavioral control, as well as describe how they are interdependent at times, and mutually interfering at others. We will establish the role that predictions play in guiding these different behavioral systems, and we will consider how these systems differ in how they develop their predictions. Finally we will also evaluate the possibility that an additional 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 more fruitful dichotomies to which we may graft the brain's varied control strategies is whether behavior is motivated by the onset of a stimulus, or whether it's directed toward a goal outcome. Historically, stimulus-driven responses (Thorndike, 1898), so named to conjure images of inanimate objects slave to the laws of physics, have been contrasted with goal-directed actions deliberatively dispatched to achieve a goal (Tolman, 1948). Arguments as to whether human (and animal) behavior is ruled by one strategy or the other have been calmed by theory and evidence suggesting that both types of behavioral control coexist. We begin here by outlining 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. Since these behaviors are instigated by a particular

stimulus or class of stimuli, they offer the advantages of being 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, which implies that they do not depend on synaptic plasticity and are often implemented at the level of the spinal cord and brainstem (Thibodeau, et al. 1992). It is generally thought that reflexes have been acquired over the course of evolution, as they are present in the simplest organisms such as bacteria all the way up to humans, and because analogous motor reflexes to the same stimulus are present across species. Example reflexes are the withdrawal reflex that comes from touching a hot surface, the startle response that is elicited in response to sudden stimuli, or 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 succeeding escape responses, and the salivary response helps with the consumption and digestion of food.

Reflexes are fundamentally reactive in that an unanticipated triggering stimulus elicits a pre-programmed response. However, significant advantage comes with being able to issue responses in a prospective manner, in anticipation of an event that requires a response. 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 or non-instrumental 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 with reflexes, Pavlovian learning is present in many invertebrates, including insects such as drosophila (Tully and Quinn, 1985), and even in the sea-slug (aplysia; Walters et al., 1981), but also in vertebrates, including humans (Davey, 1992).

The type of behavioral response emitted 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, while 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 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 achieve this, many organisms are also equipped with a mechanism through which they can learn to perform specific yet arbitrary behavioral responses (such as a lever-press) in a specific context. Here, 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 reinforcement (i.e. avoidance of an aversive outcome). While this strategy affords significant benefits in terms of cognitive efficiency, speed and accuracy, this comes at a cost. Critically, based on evidence we will review shortly, the execution of this class of behavior does not involve an anticipation of a particular outcome (Thorndike, 1898), which means behavior can become habitual, so that it cannot be flexibly adjusted should outcome

valuation suddenly change. Thus, to the organism's potential detriment, habits may persist even if their outcomes are no longer beneficial, which is thought to give rise to various forms of addiction (Everitt & Robbins, 2016).

Goal-directed control

Goal-directed control refers to a class of 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, providing the ability to adapt to changes in the environment without having to resample environmental contingencies directly. However, interrogating a cognitive map in order to generate a behavioral plan also implies that goal-directed control is cognitively demanding and slow.

Goal-directed control has been distinguished from habitual stimulus-driven behavior experimentally by training an animal to perform unique actions (e.g. lever-press or chain pull) 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

While it has been shown that rats are capable of performing in a goal-directed manner (Dickinson & Balleine, 1994), it has also been shown that those same animals may also exhibit habitual tendencies (Dickinson et al., 1995). After animals were exposed to extensive training, for example on variable interval schedules of reward (Dickinson et al., 1983), they were found to persistently elicit responses associated with devalued outcomes. 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. 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 over-training manipulation to that performed in rodents, humans also exhibit reduced outcome sensitivity consistent with the behavioral expression of habit (Tricomi et al., 2009).

Even though the distinction between habitual and goal-directed control is often drawn and investigated within the context of instrumental behavior, there is tentative evidence that a similar distinction can be drawn for Pavlovian behavior. Critically, the core criterion to distinguish habitual from goal-directed behavior is also present for conditioned Pavlovian responses, as some are more sensitive (Dayan & Berridge, 2014) than others to outcome value (Nasser et al., 2015). Interestingly, Pavlovian conditioned responses are nevertheless often considered to be stimulus-driven in a manner analogous to habits in the instrumental domain, epitomized by the prevalent assumption that incremental synaptic plasticity implements the acquisition of Pavlovian contingencies (Rescorla & Wagner, 1972). However, this form stimulus-driven Pavlovian conditioning could not account for findings showing altered patterns in the conditioned response immediately after devaluation, prior to any resampling of the environment's

contingencies (Dayan and Berridge, 2014). Despite the fact that there is evidence for the existence of these two distinct strategies within Pavlovian learning, the majority of the research in this domain has been performed using instrumental conditioning, and we focus on this theme in the remainder of this review as well, although we will revisit the Pavlovian case later.

Why multiple systems?

Given that these various strategies seem to be present in a range of organisms, why do they exist simultaneously? In other words, why are human behaviors often driven by a habitual stimulus-response system, when we have the machinery for more flexible goal-directed actions instead? One explanation could be that these behavioral control systems coexist because evolutionary adaptation occurred incrementally. These adaptations may simply have occurred through the addition of new brain circuitry without refurbishing or repurposing 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 multi-controller strategy in the absence of some additional benefit.

A second, more compelling explanation for the coexistence of multiple behavioral control systems might be that the brain's control systems share mutually beneficial interdependencies. Evolutionary 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 evolutionary more 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, while the strategic guidance of advanced systems helps elemental 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 that comes with 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. This strategy may offer significant advantages to a predator stalking its prey, but could 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 (not a rich cognitive map).

In sum, 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 in order 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. Padoa-Schioppa et al., 2006; Platt & Glimcher, 1999; Rangel et al., 2008; Balleine et al., 2008; Glimcher et al., 2013; Camerer et al., 2005) but not all (see Strait et al., 2014; Gigerenzer et al., 2011) theories of value-based decision-making, as applied to the brain, is that in order 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. Doing so facilitates a comparative process, allowing the agent to identify and pursue the option leading to the greatest expected value. This idea has motivated a search for neural representations of value predictions in the brain, an endeavor that has been enormously fruitful (but see also O'Doherty (2014) for some caveats). Accordingly, 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 striatum, as well as in a number of other brain areas such as parietal, premotor and dorsal frontal areas. We will revisit these signals in more detail later on in this review.

Reinforcement-learning

Evidence for value signals in the brain opens up the question of how such signals could be learned or acquired in the first place. Insight into a potential mechanism for this has come from the seminal work of Wolfram Schultz and colleagues, who found that the phasic activity of dopamine neurons encodes a prediction error, which signals the difference between expected and actual rewards (Schultz et al., 1997). 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 (Sutton, 1988; Montague et al., 1996). 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 motivated specific proposals that RPE signals carried by phasic dopamine facilitate neural plasticity associated with the acquisition of value predictions in these target areas, which could be applied to learn associations among stimuli, responses, or even cognitive states.

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 others (2005) proposed that the distinction between stimulus-driven 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 by virtue of integrating across past reinforcement alone. 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 doesn't depend on a model of the environment) gives rise to value representation that resembles stimulus-based association.

In order to account for goal-directed control, Daw and colleagues proposed that the agent encodes an internal model of the decision problem, consisting of the relevant states, 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 these expected values with 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 (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. We can therefore ask how these are implemented in the brain as means of moving toward a better understanding of the brain's computational composition. Here, we outline some of the key representations and signals associated with various forms of RL, and their neural correlates. See Figure 1 for an illustration of the main brain regions and functions discussed in this section.

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 (O'Keefe & Dostrovsky, 1971). Activity in these cells can represent the animal's trajectory during a spatial decision-making task, consistent with a role for place cell representations 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, such as in encoding relationships between stimuli and outcomes, the identity and category membership information about objects (Eichenbaum et al., 1999), and even maps of social hierarchy in humans (Tavares et al., 2015).

Although there is evidence 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 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 OFC in response to stimuli predictive of those outcomes (Stalnaker et al., 2009). This may provide a mechanism through which the expected value of a particular stimulus or state could be computed. While still a matter of debate, the bulk of evidence suggests that 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.

Yet, goal-directed action selection demands some form of action representation, and state transitions afforded by performing actions. A large literature has implicated an important role for

regions of posterior parietal cortex such as the lateral intraparietal sulcus in perceptual decision-making, a critical aspect of state identification (Shadlen & Newsome, 2001). Notably, neurons in posterior parietal cortex have been implicated in encoding information about stimulus category membership, which could be important for establishing current and future potential states (Freedman & Assad, 2006). Indeed, recent work has shown that the category of a prospective stimulus appears to engage these regions of the brain (Doll et al., 2015). Critically, neurons in posterior parietal cortex are implicated in encoding associations between arbitrary stimuli, which indicate the implementation of specific actions (Dorris & Glimcher, 2004). A region of 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 action contingencies (Liljeholm et al., 2011, 2013). Together, these findings suggest a role for 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 did actually occur. This SPE can then be used to adjust state transition expectations. In essence, SPEs are similar to RPEs except that SPEs are not used to learn about reward expectation; rather, they are used to learn state expectations. Gläscher et al. (2010) used fMRI to find evidence for SPEs in posterior parietal cortex and dorsolateral prefrontal cortex while participants learned a two-step Markov Decision-Problem. 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 when reward, and therefore RPEs, were also present. SPEs in posterior parietal cortex and dorsolateral prefrontal cortex are therefore a candidate signal for 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. One possibility is that 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, while decision problems that involve selection among possible motor actions, will depend to a greater extent on action codes in posterior parietal cortex. However, precisely how these various maps might be leveraged by the brain in support of model-based learning and control still needs to be determined.

Outcome valuation during decision-making

In order 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 both the amygdala and OFC related to conditioned stimuli associated with both 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 et al., 2007, 2010; Schoenbaum et al., 1998). Furthermore, human imaging studies have revealed responses in 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 (Tobler et al., 2006; O'Doherty et al., 2002; Gottfried et al., 2002, 2003).

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 role for this structure in triggering conditioned skeletomotor reflexes, such as approach and avoidance behavior, as well as consummatory responses. We will discuss below how 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. But first, we will explore in more detail the representations and signals carried by some of these areas.

Value signals have been found in both OFC and vmPFC. Electrophysiological recordings in area 13 of central OFC of non-human primates, revealed that neurons in this area encode the value of the 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, while 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 (Schoenbaum et al., 1998; McDannald et al,. 2011). Other monkey neurophysiology studies have reported neuronal responses correlating with the value of prospective outcomes throughout OFC and in other brain regions including lateral prefrontal and anterior cingulate cortex (Wallis & Miller, 2003; Lee et al., 2007; Seo et al., 2007; Smith et al., 2010). Interestingly, neurons in lateral prefrontal cortex have been found to respond in line 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). Similar representations seem to be encoded in vmPFC of humans. Activity in 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, that 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).

Not only are organisms forced to choose among rewards of varying probability and magnitude, but they are also forced to choose among rewards that differ in type. One strategy of coping with this issue is to represent and compare outcome values in a common currency. Indeed, activity in overlapping regions of vmPFC correlated with the subjective value of three distinct categories of goods in a WTP task: food items, non-food consumer items and money (Chib et al., 2009). Levy and Glimcher (2012) found evidence for a common currency in vmPFC by giving people explicit choices between different types of goods, specifically money vs. 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 are also consistent with the possibility that overlapping regions at the group level are produced by non-overlapping value representations at the voxel level of individual subjects. Using a similar paradigm as 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 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, consistent with the idea

that these regions represent value in a preliminary category specific form prior to being converted into a common currency in more dorsal parts of vmPFC. Interestingly, no region was found to uniquely encode the distributed value of monetary items, which were only found to be represented in 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 vmPFC in which the value of various outcomes are proportionally scaled in accordance with subjective value irrespective of the category from which it is drawn. Next we will consider how other information relevant for model-based computations are 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 OFC in responding to experienced outcomes, including monetary rewards (O'Doherty et al., 2001; Knutson et al., 2001; Smith et al., 2010), taste, odor and flavor (de Araujo et al., 2003b; Rolls et al., 2003), attractive faces and the aesthetic value of abstract art (O'Doherty et al., 2003; Kirk et al., 2009). These outcome representations are also strongly influenced by changes in underlying motivational states. These areas show decreasing response to food or odor or even water outcomes as motivational states change from hungry or thirsty to satiated, paralleling changes in the subjective pleasantness of the stimulus (O'Doherty et al., 2000; Small et al., 2001; de Araujo et al., 2003b; Rolls et al., 2003). 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 merely the 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 the organism has determined the value of different outcomes, it is often necessary to determine the value of available actions, based on how likely they are expected to lead to a desired outcome. In order to calculate these so-called model-based action values, a decision-making agent has to 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. without any consideration of state transitions or which outcome might be achieved, will be discussed later.

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 lateral prefrontal cortex (Miller & Cohen, 2001), as well as parts of parietal cortex implicated in numerical cognition (Platt & Glimcher, 1999), are likely to be involved. It therefore seems reasonable to hypothesize that regions of frontal cortex and parietal cortex will play a fundamental role in the computation of model-based action-values. At least partly consistent with this possibility, Simon & Daw (2011) reported increasing activity in dorsolateral prefrontal cortex and anterior cingulate cortex as a function of the depth of model-based planning during a spatial navigation task. In addition, areas of posterior parietal cortex are also known to be 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, as well as representing the target state of the action trajectories in both monkeys (MacKay, 1992; Cohen & Andersen 2002; Andersen et al. 1997) 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 prelimbic cortex and targets in basal ganglia, namely dorsomedial striatum, 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 (Ragozzino et al., 2002; Yin et al., 2005; Balleine & Dickinson, 1998; Baker & Ragozzino, 2014). While 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). On the other hand, the dorsomedial striatum appears to be necessary for both acquisition and expression of goal-directed behavior (Yin et al., 2005).

The rodent prelimbic cortex and dorsomedial striatum have been argued to correspond to the primate vmPFC and caudate nucleus, respectively (Balleine & O'Doherty, 2009). Indeed, in addition to representing the value of different outcomes on offer (discussed in previous section), activity in vmPFC also tracks instrumental contingencies, 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 (Matsumoto et al., 2003; Liljeholm et al., 2011). Contingency manipulations have also implicated the caudate nucleus in goal-directed behavior in non-human primates (Hikosaka et al., 1989) and humans (Liljeholm et al., 2011). Furthermore, activity in vmPFC has been found to track the current incentive value of an instrumental action, so that following devaluation, activity decreases for an action associated with a devalued outcome relative to an action associated with a still valued action (Valentin et al., 2007; de Wit et al., 2009). Interestingly, the strength of the connection between 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 they can be compared at decision points. While a number of studies have reported evidence for pre-choice action-values, few studies have distinguished 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 dorsal cortex, including parietal and supplementary motor cortices (Platt & Glimcher, 1999; Samejima et al., 2005; Lau & Glimcher, 2008; Sohn & Lee, 2007; Kolb et al. 1994; Whitlock et al. 2012; Wilber et al. 2014). Human fMRI studies report evidence that putative action-value signals are present in dorsal cortex,, including the supplementary motor cortex, lateral parietal and dorsolateral cortex (Wunderlich et al., 2009; Hare et al., 2011; Morris et al., 2014).

Little is known about how the range of variables that appear to influence action selection are integrated. One candidate region is the dorsomedial prefrontal cortex. In monkeys, Hosokawa and colleagues 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 (Hosokawa et al., 2013). Hunt et al. (2014) also implicated a region of 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 in order to compute an overall action-value.

Neurobiological substrates of model-free action-selection

As briefly reviewed earlier, 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). There is evidence for a role of reward-related prediction errors in learning in humans as well. A large number of 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 (O'Doherty et al., 2003, 2004; Wittmann et al., 2005; D'Ardenne et al., 2008; Pauli et al., 2015).

There is evidence to suggest 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 activity in this area correlated with the value of over-trained 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 posterior putamen (Horga et al., 2015).

The phasic activity of dopamine neurons is causally related to the instrumental actions learning via dopamine-modulated plasticity in target areas of these neurons, such as the dorsolateral striatum (Faure et al., 2005; Schoenbaum et al., 2013; Steinberg et al., 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 (Schonberg et al., 2007), while the administration of drugs that modulate dopamine function such as L-DOPA or dopaminergic antagonists influence 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 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 right posterolateral striatum and premotor cortex across individuals is associated with the degree to which individuals show evidence of habitual behavior on a task in which goal-directed and habitual responding are put in conflict with each other (De Wit et al., 2012).

Other decision variables: effort and uncertainty

One variable that is likely to play an important role during decision-making is the amount of effort, be it 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 where 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 insular cortex (Prévost et al.,

2010). Additional studies in rodents point toward the anterior cingulate cortex as playing a critical role in effortful behavior (Walton et al., 2009; Hillman & Bilkey, 2012).

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 as 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; Paulus et al., 2003; Huettel et al., 2006; 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, but 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 tradeoff 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 frontopolar cortex (Yoshida & Ishii, 2006; Daw et al., 2006, Badre et al., 2012).

Model-free and model-based Pavlovian learning

Next we turn our attention to the computations that underpin acquisition and expression of Pavlovian conditioned responses. As described previously, model-free RL has been proposed as a mechanism to underpin learning at least in 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. Yet it is known that many conditioned Pavlovian responses are strongly devaluation sensitive (Dayan & Berridge, 2014). This has led to suggestions that model-based learning mechanisms might also apply in the case of Pavlovian conditioning (Prévost et al. 2013; Dayan & Berridge, 2014).

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. In essence, one might expect the mechanism for model-based Pavlovian conditioning to be similar to that involved in model-based instrumental control, except that there is no need for the model to represent action contingencies. 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 is sensory preconditioning. Here, two cues are repeatedly paired together in the absence of reward. Following this, one of the cues is paired with reward. It was found that the cue which had not been paired with reward also spontaneously elicit appetitive conditioned responses (Rescorla, 1980).

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

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

Two distinct forms of Pavlovian appetitive conditioning can be distinguished in rodents: sign-tracking and goal-tracking (Jenkins & Moore 1973; Hearst & Jenkins 1974; Boakes 1977). Sign-tracking 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 behavioral 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 has 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 to 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 to the cue did not have immediate access to the 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 as opposed to model-based learning.

Interaction among behavioral control systems

Having considered evidence regarding the existence of multiple control systems in the brain, as well as reviewing ideas and emerging evidence about the possible neural computations underpinning each of these systems, we will briefly consider how these systems interact. There is evidence to suggest that stimulus-driven, goal-directed, and non-instrumental 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), recent work has begun to probe whether the brain leverages its varied control systems to do the same (Botvinick et al., 2009; Botvinick 2012). Evidence from human fMRI studies show that higher levels of abstraction

progressively engage more anterior regions of frontal cortex, suggesting a hierarchical organization of abstraction along a rostral-caudal axis (Koechlin et al., 2003; Badre & D'Esposito, 2007; Donoso et al., 2014). Other studies have reported signals consistent with hierarchical event structuring (Schapiro et al., 2013) and prediction errors (Ribas-Fernandes et al., 2011; Diuk et al., 2013). Although the most common depiction of hierarchical control positions the stimulus-driven system as subservient to the goal-directed system (Dezfouli & Balleine 2013), recent 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. A canonical example of an adaptive interaction between systems could be situations in which control is assigned to the goal-directed system in the early stages of behavioral acquisition. Once the problem space has been sufficiently sampled, behavioral control transitions to the habit 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 in learning, the stimulus-driven 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 habit system, it may guide the agent toward an unfavorable course of action under circumstances where 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). There are numerous examples of maladaptive interactions in the realm of psychiatric disease. For instance, habits may persist for a drug of abuse, even if the goal of the individual is to stop taking the drug (Everitt & Robbins, 2016). Overeating or compulsive behaviors may also be examples where the habit system exerts inappropriate and ultimately detrimental control over behavior (Voon et al., 2014). 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 across individuals 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 (or PIT) (Lovibond, 1983). PIT effects are typically manifested as increased instrumental response vigor in the presence of a reward predicting Pavlovian conditioned stimulus (Estes, 1943). A distinction can be made between general and specific PIT. General PIT refers to circumstances where 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 (Rescorla & Solomon, 1967; Holland & Gallagher, 2003; Corbit & Balleine, 2005).

A normative relationship between incentives and instrumental response is that 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 where PIT effects continue to exert an energizing effect on instrumental actions associated with a devalued outcome (Holland, 2004; Watson et al., 2014; though see Allman et al., 2010). This suggests that PIT effects selectively

involve the stimulus-directed 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”, which has been linked to dopaminergic regions of midbrain (Mobbs et al., 2009; Zedelius et al. 2011; Chib et al., 2014). 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 under a situation where the motivation to succeed is very high. A number of theories have been proposed to account for choking effects, reflecting various 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, under a situation where the habitual system is better placed to reliably execute a skilled behavior. While 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). It is possible that one or more of these ideas could hold true in that 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. One influential hypothesis has been that there exists an arbitrator that determines the influence each system has over behavior based on a number of criteria. One important factor is the relative predictive accuracy over which action should be selected within the two systems such that, all else being equal, behavior should be controlled by the system with the most accurate prediction (Daw et al., 2005). Leveraging 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 which assigns behavioral control as a function of system reliability. Connectivity between the arbitration areas and regions of the brain encoding habitual but not goal-directed action-values were 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 an goal-directed policy is warranted, then perhaps it achieves this through active inhibition of the habitual system, thereby 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) or 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. It is known that changes in cognitive strategies or appraisal implemented via prefrontal cortex can influence the likelihood of both aversive and appetitive Pavlovian conditioned responses, perhaps via down-regulation of the amygdala and ventral striatum (Delgado et al., 2008a, 2008b; Staudinger et al., 2009). This type of “top-down” process could be viewed as a form of arbitration, in which Pavlovian control policies are down-weighted in situations where 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 where 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 not be appropriate or relevant. A commensurate prediction is that any such arbitration process would happen at a relatively slower time scale relative to the more rapid response time available to the Pavlovian system. Therefore, traces of initial Pavlovian control might become manifest in behavior even under situations where the arbitration system subsequently implements an inhibition of the Pavlovian system.

Neural systems for learning and inference in a social context

Until now 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, other 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 the current review. However, one question we can briefly consider is whether value-based action selection in social contexts depends on similar or distinct control systems and neural circuitry as are involved in value-based action selection in non-social contexts.

One of the simplest extensions of the framework we have discussed so far to the social domain is to the mechanisms underlying observational learning where an agent can learn about the value of stimuli or actions not through direct experience but instead through observing the behavior of another agent. A number of studies have revealed engagement of brain regions including the ventral and dorsal striatum and 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 where it is necessary to learn from the actions being taken by others the brain may rely on circuitry and similar updating signals as known to be involved in model-based RL (Seo et al., 2009; Abe & Lee, 2011; Liljeholm et al., 2012).

However, under some social situations, additional circuitry that has been implicated in mentalizing or theory of mind may be engaged (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 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 is 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 anterior medial prefrontal cortex, as well as in a region of temporal-parietal junction. Similarly, Boorman et al., (2013) found evidence for updating signals related to learning about another individual's expertise on a financial investment task in temporal-parietal junction and dorsomedial frontal cortex. Suzuki et al., (2015) found evidence for the representation of beliefs in posterior superior

temporal sulcus about the likely future actions of a group of individuals, and moreover found that this activity was specifically engaged when performing in social as compared to a non-social context.

Taken together, these findings suggest that while learning and making decisions in a social context often depends on overlapping brain circuitry as when learning in non-social contexts, on occasion where it is necessary to know about relevant features of another agent, additional distinct circuitry is deployed in order facilitate socially relevant tasks such as inferring the internal mental states of others.

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 made considerable headway not only in terms of understanding the algorithms embodied by the brain, but also helped to illuminate how the brain navigates the tradeoffs 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 showing 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.

On the back of these advances, new unresolved issues have emerged. We have reviewed evidence from both animal and human studies that a goal-directed (model-based) system guides behavior in some circumstances, but other situations favor a habitual (model-free) strategy. Factors such as task familiarity, task complexity, and reward contingencies may influence the tradeoff 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 (“Do significant benefits come from favoring one strategy over another?”), cognitive capacities (“Is the brain ‘aware’ of its own limitations?”), and social context may play a role in system deployment. Also unknown is whether Pavlovian drives factor into the arbitration scheme used to determine behavioral control.

Furthermore, we understand very 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 this can be interrupted by a more flexible goal-directed strategy if needed. However, this raises the question of what the model-based system is doing when it’s not favored for control: Is the model-based system passively working the background, waiting to be called back into the game, or is it moved offline to conserve resources. If so, how is it brought back online in a sensible way? We must also ask the question of 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; that is, 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 tradeoff. Almost nothing is known about the role played by the brain's varied control system with respect to exploration. Given the exploitive advantages that come with having multiple control strategies at one's disposal, some of which we have outlined here, are there 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 vmPFC and anterior cingulate cortex do reflect knowledge of strategic information and appears to receive this information via inputs from the mentalizing network (Hampton et al., 2008; Suzuki et al. 2015). Whether these mentalizing-related computations can be considered to be a "fourth" system for guiding behavior, or instead can be considered 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 gets engaged in a particular situation is currently unknown, although it 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 (Montague et al. 2012; Maia & Frank 2011; Huys et al. 2016). 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 mechanisms through which they manifest, novel treatments may present themselves (be it behavioral, chemical, or mechanistic) to help millions of people lead more fulfilling lives.

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.

Acknowledgements

This work was supported by an NIH conte center grant on the neurobiology of social decision making (P50MH094258-01A1), NIH grant DA033077-01 (supported by OppNet, NIH's Basic Behavioral and Social Science Opportunity Network) and National Science Foundation grant 1207573 to JOD.

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

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

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

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 Research*. 238(2):457–62

Ariely D, Gneezy U, Loewenstein G, Mazar N. 2009. Large Stakes and Big Mistakes. *Review of Economic Studies*. 76(2):451–69

Badre D, D'Esposito M. 2007. Functional Magnetic Resonance Imaging Evidence for a Hierarchical Organization of the Prefrontal Cortex. *Journal of Cognitive Neuroscience*. 19(12):2082–99

Badre D, Doll BB, Long NM, Frank MJ. 2012. Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration. *Neuron*. 73(3):595–607

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

Balleine B, Dickinson A. 1991. Instrumental performance following reinforcer devaluation depends upon incentive learning. *The Quarterly Journal of Experimental Psychology Section B*. 43(3):279–96

Balleine BW, Daw ND, O'Doherty JP. 2008. Multiple forms of value learning and the function of dopamine. *Neuroeconomics: decision making and the brain*. 36:7–385

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

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

Boakes RA. 1977. Performance on learning to associate a stimulus with positive reinforcement. *Operant-Pavlovian interactions*. 67–97

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

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

Botvinick MM. 2012. Hierarchical RL and decision making. *Current Opinion in Neurobiology*. 22(6):956–62

Botvinick MM, Niv Y, Barto AC. 2009. Hierarchically organized behavior and its neural foundations: A RL perspective. *Cognition*. 113(3):262–80

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

Camerer C, Loewenstein G, Prelec D. 2005. Neuroeconomics: How neuroscience can inform economics. *Journal of economic Literature*. 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

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

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
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
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
Corbit LH, Balleine BW. 2000. The Role of the Hippocampus in Instrumental Conditioning. *J. Neurosci.* 20(11):4233–39
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
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
Cushman F, Morris A. 2015. Habitual control of goal selection in humans. *Proc Natl Acad Sci U S A*. 112(45):13817–22
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
Davey GCL. 1992. Classical conditioning and the acquisition of human fears and phobias: A review and synthesis of the literature. *Advances in Behaviour Research and Therapy*. 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
Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. 2006. Cortical substrates for exploratory decisions in humans. *Nature*. 441(7095):876–79
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
De Araujo IET, Kringelbach ML, Rolls ET, McGlone F. 2003a. Human Cortical Responses to Water in the Mouth, and the Effects of Thirst. *Journal of Neurophysiology*. 90(3):1865–76
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. *European Journal of Neuroscience*. 18(7):2059–68
De Araujo IET, Rolls ET, Velazco MI, Margot C, Cayeux I. 2005. Cognitive Modulation of Olfactory Processing. *Neuron*. 46(4):671–79
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):3787–3800
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
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
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):
Dickinson A. 1985. Actions and Habits: The Development of Behavioural Autonomy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 308(1135):67–78
Dickinson A, Balleine B. 1994. Motivational control of goal-directed action. *Animal Learning & Behavior*. 22(1):1–18

Dickinson A, Balleine B, Watt A, Gonzalez F, Boakes RA. 1995. Motivational control after extended instrumental training. *Animal Learning & Behavior*. 23(2):197–206

Dickinson A, Nicholas DJ, Adams CD. 1983. The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section 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):5797–5805

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

Doll BB, Hutchison KE, Frank MJ. 2011. Dopaminergic Genes Predict Individual Differences in Susceptibility to Confirmation Bias. *J. Neurosci.* 31(16):6188–98

Dorris MC, Glimcher PW. 2004. Activity in Posterior Parietal Cortex Is Correlated with the Relative Subjective Desirability of Action. *Neuron*. 44(2):365–78

Doya K. 1999. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*. 12(7–8):961–74

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

Estes WK. 1943. Discriminative conditioning. I. A discriminative property of conditioned anticipation. *Journal of Experimental Psychology*. 32(2):150–55

Everitt BJ, Robbins TW. 2016. Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annual Review of Psychology*. 67(1):23–50

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

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

Flagel SB, Watson SJ, Robinson TE, Akil H. 2007. Individual differences in the propensity to approach signals vs goals promote different adaptations in the dopamine system of rats. *Psychopharmacology (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

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

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

Gigerenzer G, Gaissmaier W. 2011. Heuristic Decision Making. *Annual Review of Psychology*. 62(1):451–82

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

Glimcher PW, Fehr E. 2013. *Neuroeconomics: Decision Making and the Brain*. Academic Press

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

Gottfried JA, O'Doherty J, Dolan RJ. 2003. Encoding Predictive Reward Value in Human Amygdala and OFC. *Science*. 301(5636):1104–7

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*, eds. AM Thierry, J Glowinski, PS Goldman-Rakic, Y Christen, pp. 51–77. Springer Berlin Heidelberg

Hampton AN, Bossaerts P, O'Doherty JP. 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl. Acad. Sci. U.S.A.* 105(18):6741–46

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

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

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

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

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

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. *European Journal of Neuroscience*. 17(8):1680–94

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

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

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

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

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

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

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

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

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

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

Koechlin E, Ody C, Kouneiher F. 2003. The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*. 302(5648):1181–85

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

Konorski J, Miller S. 1937. On Two Types of Conditioned Reflex. *The Journal of General Psychology*. 16(1):264–72

Kuvayev L, Sutton R. 1996. Model-Based RL with an Approximate, Learned Model. *IN PROCEEDINGS OF THE NINTH YALE WORKSHOP ON ADAPTIVE AND LEARNING SYSTEMS*, pp. 101–5

Lau B, Glimcher PW. 2008. Value Representations in the Primate Striatum during Matching Behavior. *Neuron*. 58(3):451–63

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

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
Lee SW, Shimojo S, O'Doherty JP. 2014. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*. 81(3):687–99
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
Levy DJ, Glimcher PW. 2012. The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*. 22(6):1027–38
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
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
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
Lovibond PF. 1983. Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *J Exp Psychol Anim Behav Process*. 9(3):225–47
MacKay WA. 1992. Properties of reach-related neuronal activity in cortical area 7A. *Journal of Neurophysiology*. 67(5):1335–45
Maia TV, Frank MJ. 2011. From RL models to psychiatric and neurological disorders. *Nat Neurosci*. 14(2):154–62
Matsumoto K, Suzuki W, Tanaka K. 2003. Neuronal Correlates of Goal-Based Motor Selection in the Prefrontal Cortex. *Science*. 301(5630):229–32
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):2700–2705
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
Miller EK, Cohen JD. 2001. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*. 24(1):167–202
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. *Psychological Science*. 20(8):955–62
Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* 16(5):1936–47
Montague PR, Dolan RJ, Friston KJ, Dayan P. 2012. Computational psychiatry. *Trends Cogn. Sci. (Regul. Ed.)*. 16(1):72–80
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
Nasser HM, Chen Y-W, Fiscella K, Calu DJ. 2015. Individual variability in behavioral flexibility predicts sign-tracking tendency. *Front Behav Neurosci*. 9:
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
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
O'Doherty JP. 2004. Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*. 14(6):769–76

O'Doherty JP. 2014. The problem with value. *Neuroscience & Biobehavioral Reviews*. 43:259–68
O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. 2003. Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*. 38(2):329–37
O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ. 2002. Neural Responses during Anticipation of a Primary Taste Reward. *Neuron*. 33(5):815–26
O'Keefe J, Dostrovsky J. 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*. 34(1):171–75
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
Padoa-Schioppa C, Assad JA. 2006. Neurons in the OFC encode economic value. *Nature*. 441(7090):223–26
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
Pascoe JP, Kapp BS. 1985. Electrophysiological characteristics of amygdaloid central nucleus neurons during Pavlovian fear conditioning in the rabbit. *Behavioural Brain Research*. 16(2–3):117–33
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
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. *The Journal of Neuroscience*. 35(42):14220–33
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
Pavlov I. 1927. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex.* Oxford University Press, London
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
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:
Pfeiffer BE, Foster DJ. 2013. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*. 497(7447):74–79
Plassmann H, O'Doherty J, Rangel A. 2007. OFC Encodes Willingness to Pay in Everyday Economic Transactions. *J. Neurosci.* 27(37):9984–88
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
Plassmann H, O'Doherty J, Shiv B, Rangel A. 2008. Marketing actions can modulate neural representations of experienced pleasantness. *PNAS*. 105(3):1050–54
Platt ML, Glimcher PW. 1999. Neural correlates of decision variables in parietal cortex. *Nature*. 400(6741):233–38
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
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
Ragozzino ME, Ragozzino KE, Y J, Kesner RP. 2002. Role of the dorsomedial striatum in behavioral flexibility for response and visual cue discrimination learning. *Behavioral Neuroscience*. 116(1):105–15
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

Rangel A, Hare T. 2010. Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*. 20(2):262–70
Rescorla RA. 1980. Simultaneous and successive associations in sensory preconditioning. *J Exp Psychol Anim Behav Process*. 6(3):207–16
Rescorla RA, Solomon RL. 1967. Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychol Rev*. 74(3):151–82
Rescorla RA, Wagner AR, others. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*. 2:64–99
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
Rolls ET, Kringelbach ML, De Araujo IET. 2003. Different representations of pleasant and unpleasant odours in the human brain. *European Journal of Neuroscience*. 18(3):695–703
Salzman CD, Fusi S. 2010. Emotion, Cognition, and Mental State Representation in Amygdala and Prefrontal Cortex. *Annu Rev Neurosci*. 33:173–202
Salzman CD, Paton JJ, Belova MA, Morrison SE. 2007. Flexible Neural Representations of Value in the Primate Brain. *Annals of the New York Academy of Sciences*. 1121(1):336–54
Samejima K, Ueda Y, Doya K, Kimura M. 2005. Representation of Action-Specific Reward Values in the Striatum. *Science*. 310(5752):1337–40
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
Schoenbaum G, Chiba AA, Gallagher M. 1998. OFC and basolateral amygdala encode expected outcomes during learning. *Nat Neurosci*. 1(2):155–59
Schoenbaum G, Esber GR, Iordanova MD. 2013. Dopamine signals mimic reward prediction errors. *Nat. Neurosci.* 16(7):777–79
Schultz W, Dayan P, Montague PR. 1997. A Neural Substrate of Prediction and Reward. *Science*. 275(5306):1593–99
Seo H, Barraclough DJ, Lee D. 2009. Lateral intraparietal cortex and RL during a mixed-strategy game. *J. Neurosci.* 29(22):7278–89
Shadlen MN, Newsome WT. 2001. Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. *Journal of Neurophysiology*. 86(4):1916–36
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
Simon DA, Daw ND. 2011. Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. *J. Neurosci.* 31(14):5526–39
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
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
Sohn J-W, Lee D. 2007. Order-Dependent Modulation of Directional Signals in the Supplementary and Presupplementary Motor Areas. *J. Neurosci.* 27(50):13655–66
Stalnaker TA, Franz TM, Singh T, Schoenbaum G. 2007. Basolateral amygdala lesions abolish orbitofrontal-dependent reversal impairments. *Neuron*. 54(1):51–58
Steinberg EE, Janak PH. 2013. Establishing causality for dopamine in neural function and behavior with optogenetics. *Brain Research*. 1511:46–64
Strait CE, Blanchard TC, Hayden BY. 2014. Reward Value Comparison via Mutual Inhibition in VmPFC. *Neuron*. 82(6):1357–66
Sutton R, Precup D, Singh S. 1999. Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in RL. *Artificial Intelligence*. (112):181–211

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. *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats*, pp. 288–96. Cambridge, MA, USA: MIT Press
Suzuki S, Adachi R, Dunne S, Bossaerts P, O'Doherty JP. 2015. Neural mechanisms underlying human consensus decision-making. *Neuron*. 86(2):591–602
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
Thibodeau GA, Patton KT, Wills. 1992. *Structure & Function of the Body*. Mosby Year Book
Thorndike EL. 1898. Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*. 2(4):i – 109
Tobler PN, O'Doherty JP, Dolan RJ, Schultz W. 2006. Human Neural Learning Depends on Reward Prediction Errors in the Blocking Paradigm. *Journal of Neurophysiology*. 95(1):301–10
Tolman EC. 1948. Cognitive maps in rats and men. *Psychological review*. 55(4):189
Tricomi E, Balleine BW, O'Doherty JP. 2009. A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*. 29(11):2225–32
Tully T, Quinn WG. 1985. Classical conditioning and retention in normal and mutantDrosophila melanogaster. *J. Comp. Physiol.* 157(2):263–77
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
Wallis JD, Miller EK. 2003. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*. 18(7):2069–81
Walters ET, Carew TJ, Kandel ER. 1981. Associative Learning in Aplysia: evidence for conditioned fear in an invertebrate. *Science*. 211(4481):504–6
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. *European Journal of Neuroscience*. 29(8):1678–91
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
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
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
Wilson RC, Takahashi YK, Schoenbaum G, Niv Y. 2014. OFC as a Cognitive Map of Task Space. *Neuron*. 81(2):267–79
Wimmer GE, Shohamy D. 2012. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science*. 338(6104):270–73
Winn P, Brown VJ, Inglis WL. 1997. On the Relationships Between the Striatum and the Pedunculopontine Tegmental Nucleus. *Critical Reviews™ in Neurobiology*. 11(4):241–61
Wit S de, 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
Wit S de, Watson P, Harsay HA, Cohen MX, Vijver I van de, Ridderinkhof KR. 2012. Corticostriatal Connectivity Underlies Individual Differences in the Balance between Habitual and Goal-Directed Action Control. *J. Neurosci.* 32(35):12066–75

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

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

Yanike M, Ferrera VP. 2014. Representation of Outcome Risk and Action in the Anterior Caudate Nucleus. *J. Neurosci.* 34(9):3279–90

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. 2011. Large-scale automated synthesis of human functional neuroimaging data. *Nat Meth*. 8(8):665–70

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

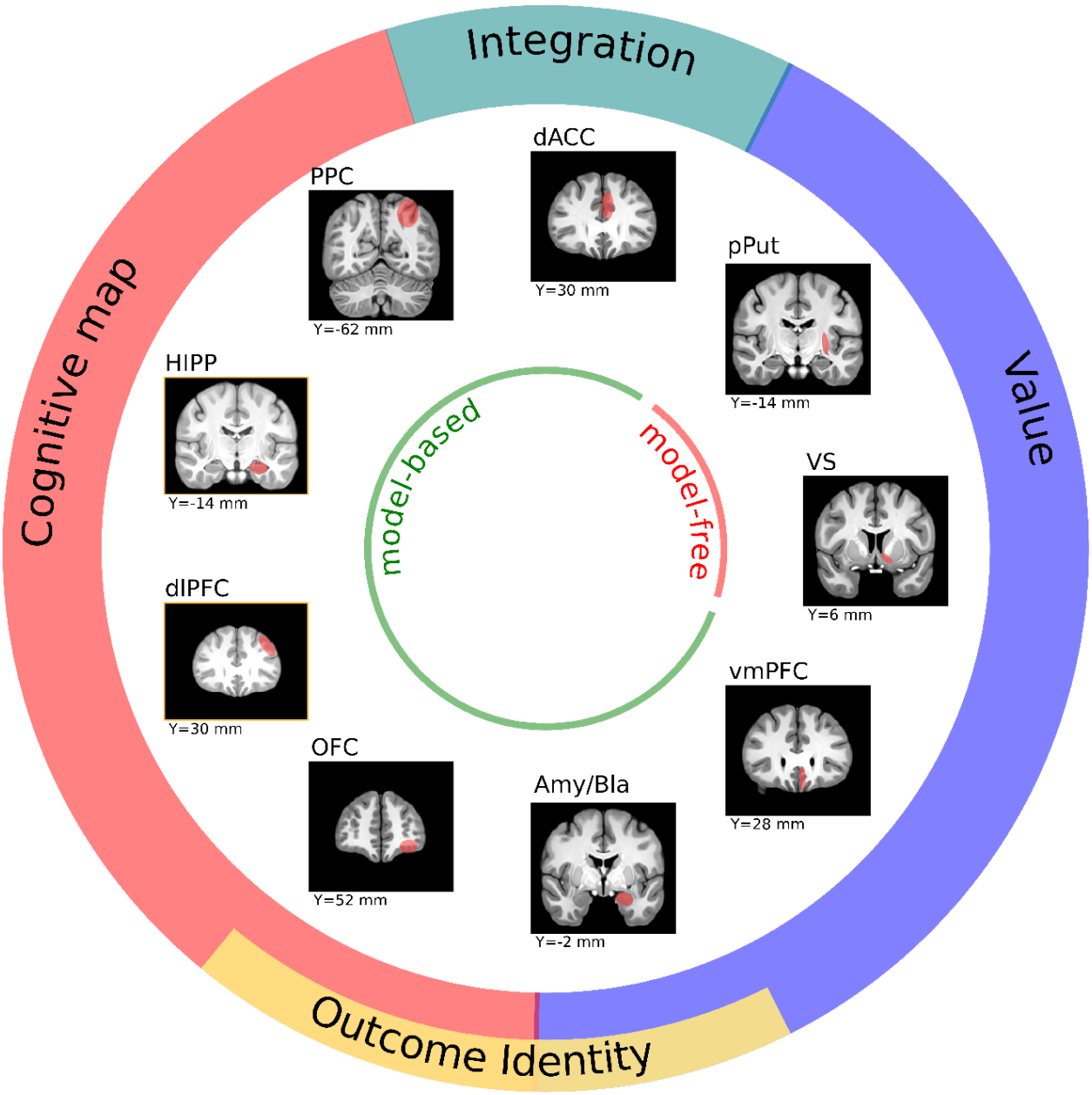
Yin HH, Knowlton BJ, Balleine BW. 2006. Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. *Behavioural Brain Research*. 166(2):189–96

Yoshida W, Ishii S. 2006. Resolution of Uncertainty in Prefrontal Cortex. *Neuron*. 50(5):781–89

Zedelius CM, Veling H, Aarts H. 2011. Boosting or choking – How conscious and unconscious reward processing modulate the active maintenance of goal-relevant information. *Consciousness and Cognition*. 20(2):355–62

Figure Captions

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, 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, while the outer circle identifies specific sub-functions 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 providing 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 locus of arbitration mechanisms are omitted from this figure for simplicity. Acronyms: Amy/BLA - basolateral complex of amygdala; dACC - dorsal cingulate cortex; dlPFC - dorsolateral prefrontal cortex; HIPP - Hippocampus; OFC – orbitofrontal cortex; pPut - posterior putamen; PPC - posterior parietal cortex; vmPFC – ventromedial prefrontal cortex; VS - ventral striatum.


Integration
Value
Outcome Identity
Cognitive map
model-based
model-free
dACC
Y=30 mm
PPC
Y=-62 mm
pPut
Y=-14 mm
HIPP
Y=-14 mm
VS
Y=6 mm
dlPFC
Y=30 mm
vmPFC
Y=28 mm
OFC
Y=52 mm
Amy/Bla
Y=-2 mm