American College of Neuropsychopharmacology

www.nature.com/npp

REVIEW ARTICLE

Reinforcement-learning in fronto-striatal circuits

Bruno Averbeck[1]✉ and John P. O'Doherty[2]✉

© The Author(s), under exclusive licence to American College of Neuropsychopharmacology 2021

We review the current state of knowledge on the computational and neural mechanisms of reinforcement-learning with a particular focus on fronto-striatal circuits. We divide the literature in this area into five broad research themes: the target of the learning—whether it be learning about the value of stimuli or about the value of actions; the nature and complexity of the algorithm used to drive the learning and inference process; how learned values get converted into choices and associated actions; the nature of state representations, and of other cognitive machinery that support the implementation of various reinforcement-learning operations. An emerging fifth area focuses on how the brain allocates or arbitrates control over different reinforcement-learning sub-systems or "experts". We will outline what is known about the role of the prefrontal cortex and striatum in implementing each of these functions. We then conclude by arguing that it will be necessary to build bridges from algorithmic level descriptions of computational reinforcement-learning to implementational level models to better understand how reinforcement-learning emerges from multiple distributed neural networks in the brain.

Neuropsychopharmacology (2022) 47:147–162; https://doi.org/10.1038/s41386-021-01108-0

INTRODUCTION

Learning to exploit the environment for resources and to avoid harm are fundamental to the success of individuals and species. These learning processes can be characterized using reinforcement-learning (RL), a theoretical framework that arose from the interface between artificial intelligence and behavioral learning theory [1, 2]. RL models provide strong constraints on behavior and the features of the environment relevant to learning and have been effective at predicting behavior and neural activity.

In RL an agent, either biological or artificial, learns the values of actions or choices. Choices are then driven by a policy, which selects actions that have the highest value in the current state. In RL, states define all of the information necessary to make a choice. States can be defined, for example, by objects in the environment as well as internal states like hunger or thirst. These choices can be stochastic, such that actions are selected probabilistically, relative to how valuable they are. The learning in RL is driven by reward prediction errors, which are the difference between the reward received, and the reward that was expected, following an action. Substantial work has linked reward prediction errors to the response of mid-brain dopamine neurons (for reviews see [3, 4]). Because of the strong projection of mid-brain dopamine neurons to the striatum, it is usually thought that action values are represented in the striatum. These values are then proposed to be updated by dopamine, following the selection of an action [5].

In this review we will focus on five separate research thrusts concerning the implementation of RL in fronto-striatal circuits. The first concerns the target of the learning process, which includes whether learning is about the value of stimuli encountered in the world or whether the learning is about the actions needed to increase an individual's expected future reward. The second concerns the complexity or form of the learning algorithm, in particular the neural implementation of various flavors of RL including model-free, model-based and hierarchical RL. A third thrust concerns identifying neural mechanisms of different components of the RL process—distinguishing brain systems involved in making value predictions from those involved in learning predictions, and from those involved in action-selection and decision-making. A fourth thrust is about the nature of the representations needed as a scaffold for RL. RL algorithms operate on state-spaces—and a major research question concerns how the brain extracts relevant information from the environment to form these state-spaces, as well as how the brain infers which state of the world an agent is in given incoming sensory data. An emerging fifth theme, concerns how the brain arbitrates between different strategies for RL, such as between model-based and model-free. Finally, we will outline outstanding questions, and highlight important future directions for research in this area. We will focus in particular on cortical and striatal mechanisms of RL and will not review the functions of dopamine neurons in RL because it is beyond the scope of the current review as well as being extensively covered elsewhere [3, 4].

FRONTAL-STRIATAL SYSTEMS

Adaptive learning is critical to survival, and therefore, RL engages a broad set of neural circuits that likely include much of the cortex, beyond early sensory and motor areas, as well as the portions of the basal ganglia to which these cortical areas project. Perhaps the most important components of the cortico-basal ganglia circuits for RL are the prefrontal cortical-striatal systems.

Frontal-striatal systems are anatomically defined circuits connecting frontal cortical areas to corresponding locations in the striatum. These systems are organized topologically such that

[1]National Institute of Mental Health, Bethesda, MD, USA. [2]Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA. ✉email: averbeckbb@mail.nih.gov; jdoherty@caltech.edu

Received: 4 March 2021 Revised: 6 July 2021 Accepted: 9 July 2021
Published online: 5 August 2021

neighboring locations in the cortex project to neighboring locations in the striatum, and regions that are further away in the cortex project to distant regions in the striatum [6, 7]. The ventral striatum (nucleus accumbens) receives inputs from ventral-medial prefrontal cortex (vmPFC, approximately Brodmann's areas 14, 25, and 32). The dorsal striatum correspondingly receives inputs from dorsal-lateral prefrontal cortex (dlPFC, Brodmann's areas 46 and 8). Unlike much neural circuitry, there are no direct reciprocal connections from the striatum back to cortex. However, the frontal-striatal circuitry does form a closed loop via subsequent descending projections that maintain the topographic organization [8]. The ventral-striatum projects to the ventral-pallidum, while the dorsal-striatum projects to the dorsal pallidum. The dorsal and ventral pallidum then send projections back to the medial-dorsal (MD) nucleus of the thalamus. The MD then sends topographically organized return projections to the vmPFC and the dlPFC, closing the loop.

Although the dorsal and ventral circuits from cortex through the basal ganglia define two poles, all of frontal cortex projects topologically into the striatum, and maintains a topological organization through the subsequent circuitry. Interestingly, dorso-medial prefrontal cortex (BA 9/24) and ventro-lateral prefrontal cortex (vlPFC; BA 47/12 l) have overlapping projections to the central striatum. The central striatum is therefore a site of convergence, where multiple prefrontal-cortical areas send overlapping projections.

It is important to note that the fundamental neural architecture that underlies RL is organized as a loop. Frontal-striatal circuitry relies fundamentally on feedback for learning. And this feedback may be delivered via the looped architecture.

1) TARGET OF THE LEARNING

Here we consider the "what" that is learned about. It is possible to distinguish between different neuroanatomical and computational implementations of RL-processes depending on the target of the learning.

LEARNING ABOUT THE VALUE OF STATES VS ACTIONS IN THE STRIATUM

In RL, a fundamental distinction is made between learning about the value of states of the world, and learning about the value of the available actions in each state of the world. Learning about the value of states and actions, is analogous to the psychological constructs of Pavlovian and instrumental conditioning respectively [9, 10]. This distinction also maps onto the ventral-dorsal circuitry through the striatum [11, 12].

A classic RL algorithm making a distinction between learning about states and actions along these lines is the "actor-critic" [9]. In this model, a critic learns and makes predictions about the value of states, and computes prediction errors as the agent transitions from one state to another. These prediction errors are used not only to update "state-values" in the critic, but also to update the "policy" in the actor. The policy in RL is the function that maps from states to actions. The actor, therefore, takes actions based on its learned policy. Interest in the applicability of the actor-critic model for understanding neural RL arose from the suggestion that the distinction between the ventral and dorsal striatum might broadly map onto the distinction between the actor and the critic in the actor-critic model [10, 12]. Consistent with this proposal, a large literature from rodents to humans implicates the dorsal striatum in instrumental learning, while the ventral striatum has by contrast been implicated in Pavlovian learning [11, 13].

Lesions of the ventral striatum in rodents or the administration of dopaminergic antagonists therein is associated with impaired acquisition and expression of Pavlovian behaviors, especially a form of cue-oriented Pavlovian conditioning called sign-tracking [14, 15]. Ventral striatal lesions in non-human primates also affect learning to associate rewards with images [16, 17], although not under all conditions [18]. Rothenhoefer et al. [19] used a task in which rewards were probabilistically associated either with images independent of their location (Fig. 1F; "What" condition), or locations independent of the image presented at that location (Fig. 1E; "Where" condition). Lesions to the ventral striatum affected learning to associate images with rewards, but not actions with rewards [19]. Neurons in the ventral striatum, specifically the nucleus accumbens core, have been found to encode cue-related information during Pavlovian conditioning [20–22]. Consistent with these rodent findings, neurons in the ventral striatum in non-human primates, as well as orbito-frontal cortex and the amygdala, which are connected in a mono-synaptic circuit, have enriched representations of chosen images and their values (Fig. 1D; [23, 24]). However, neurons across these structures have almost no representation of the locations of the images, or the actions required to obtain those images (Fig. 1C).

On the other hand, several groups have shown that the dorsal striatum (DS) [25–27], as well as dorsal-lateral prefrontal cortex [27, 28], which projects to the dorsal striatum, has an enriched representation of actions and action values (Fig. 1A). These studies have also shown that DS neurons can correlate both positively and negatively with action values [26]. Relatedly, injection of D2 dopamine antagonists into the dorsal striatum, which blocks dopamine's effects on a subset of medium spiny striatal neurons, affected learning spatial sequences of saccades, while having no effect on decision making based on perceptual inference [29]. Animals could select targets using immediately available perceptual information after D2 antagonist injections into the DS. However, they had deficits when they learned over trials using reward feedback to select the appropriate target (Fig. 1B).

A broad pattern of dissociation appears evident between ventral and dorsal striatum based on different targets of the learning process from stimuli to actions. However, the actor-critic model makes a more specific prediction by which value-predictions generated by the critic are used to train the actor via a common prediction error signal. If the ventral striatum is acting as a critic, its integrity should be necessary for learning about the value of actions as well as of stimuli. However, the finding that learning about the value of actions was unaffected by lesions of ventral striatum while learning about the value of stimuli was (Fig. 1E, F), suggests that these two processes can occur independently and in parallel, inconsistent with a literal implementation of the actor-critic theory.

What other RL mechanisms might be going on instead? Action-value learning algorithms such as Q-learning and its variants do not keep track of state-values at all, but instead update the value of actions directly using an action-based (as opposed to a state-based) prediction error [30]. Therefore, an alternative account of striatal function is that the dorsal striatum is involved in direct action learning while the ventral striatum is involved in learning about the value of stimuli independently of the functions of the dorsal striatum in action-learning.

A recent fMRI study [31] used a unique design in which value and prediction errors for state- and action- learning could be dissociated. It was found that prediction error signals throughout the ventral and dorsal striatum correlated with both an actor-critic RL model and an action-learning RL model (Fig. 2), suggesting that both mechanisms might be operating within cortico-striatal circuits simultaneously. If both direct action-learning and an actor-critic mechanism operate in parallel, this could explain why lesions of the ventral striatum can leave action-learning intact, as there are multiple redundant mechanisms for underpinning learning about the value of actions, which could in principle be unaffected by a lesion to a ventral striatal critic. Consistent with the dual contribution of both actor-critic and Q-learning strategies to human behavior, several studies have reported evidence that

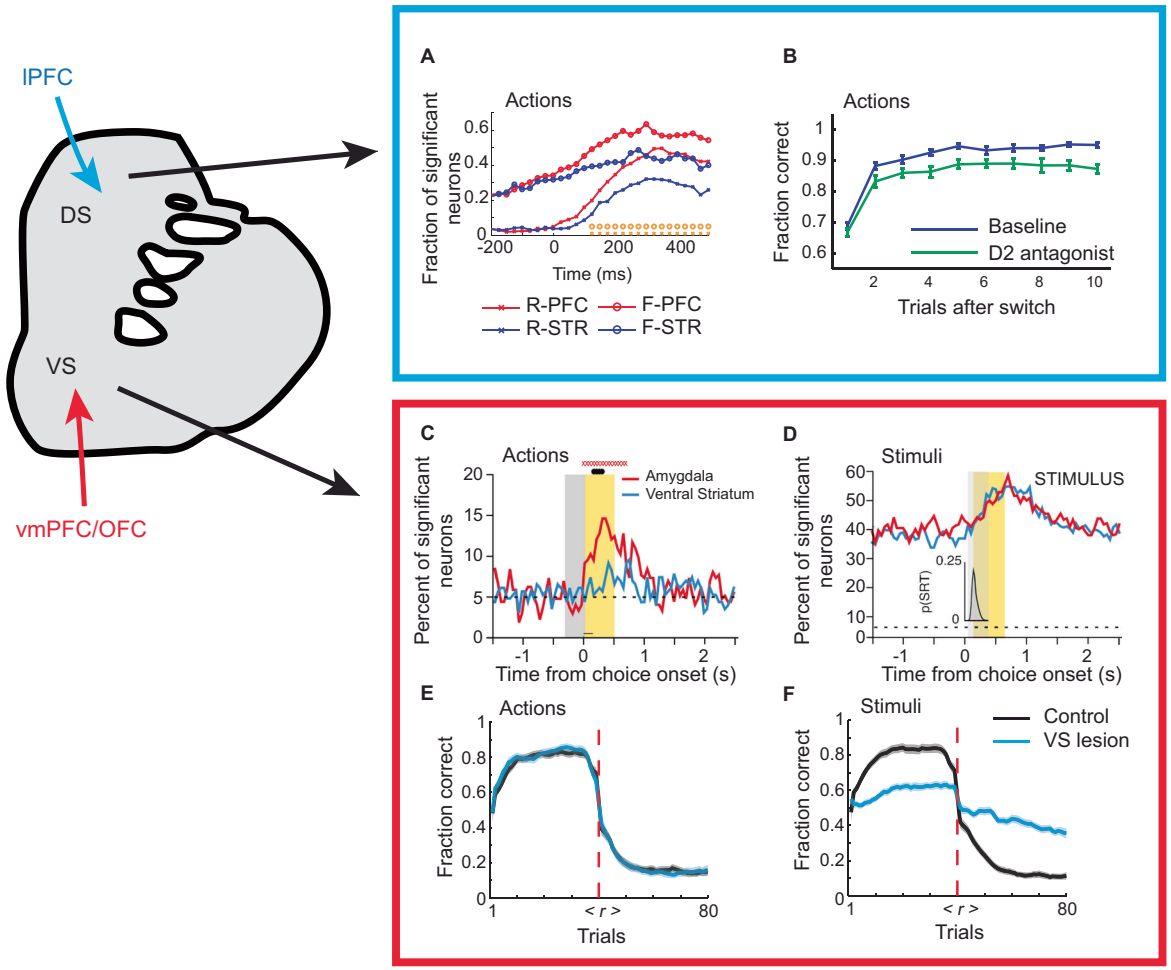

Fig. 1 Ventral-dorsal striatum and learning target. A Representation of chosen action sequences in a dorsal frontal-striatal circuit, when the spatial sequence of saccades could be learned over trials. Time 0 is the onset of the choice options. When actions can be learned (fixed condition in prefrontal cortex, F-PFC and striatum F-STR), chosen actions are represented at elevated levels before the cues that indicate choices are shown. However, when the monkeys must wait for the cues to be shown to make their choice (random condition in prefrontal cortex and striatum, R-PFC and R-STR), the representation of the choice only appears after the cues are shown. **B** Behavioral effects of injecting a dopamine D2 receptor antagonist into the dorsal striatum. The animals chose the correct action less frequently, during learning, following the D2 antagonist injection. **C** Data from a study in which the chosen object, and not the chosen action, determined the reward. Neither the amygdala or the ventral striatum have a strong representation of the chosen action. **D** The chosen stimulus, however, is strongly represented in both the amygdala and ventral striatum, when the reward is related to the stimulus. Inset shows reaction time distribution. **E** Data from an experiment in which animals had to either choose an action, independent of the stimulus to maximize reward, (panel **E**) or to choose a stimulus independent of its location to maximize reward (panel **F**). Lesions to the ventral striatum (VS lesion) have no effect on learning to choose the correct action (panel **E**) but have a large effect on learning to choose the correct stimulus (panel **F**). The dotted vertical line indicates the reversal point, where the choice-outcome mappings were reversed.

the relative degree of engagement of the actor-critic and Q-learning may vary across individuals in a manner that is related to psychopathology such as in schizophrenia [32, 33].

It should be noted that the distinction between the role of ventral striatum and dorsal striatum on stimulus vs action learning noted here does not preclude an important modulatory contribution of the ventral striatum to instrumental performance. For instance, the ventral striatum plays a role in regulating the balance between effort and reward [34, 35], as well as mediating the influence of Pavlovian cues on the vigor of instrumental responding [36, 37]. In addition, the ventral striatum has also been implicated in paradoxical decrements of reward-related motor performance that can occur in response to large incentives aka choking [38]. It should be noted that the role of motivation in regulating instrumental performance is poorly understood from a reinforcement-learning perspective (though see [39]), in part because the discrete trial-based nature of typical RL modeling applied to neuroscience does not easily translate to the free operant experiments in which motivational effects are most typically studied.

Large-scale organization of dorsal and ventral systems

The ventral vs dorsal specialization of the striatum also extends to its cortical inputs, thereby accommodating a wider distinction between the functions of ventral vs dorsal cortico-striatal networks. Within the cortex, ventral cortical areas including the orbitofrontal and ventromedial prefrontal cortex are suggested to be more involved in processing stimulus-outcome associations [40–44] and in representing features of the goal or outcome of the decision process [45, 46], while more dorsal regions including the anterior cingulate, premotor and parietal cortices are involved in action-related processing [47–50]. The notion of a brain-wide ventral vs dorsal distinction was recently elaborated on by Averbeck and Murray [51]. While the actor-critic model suggests that the ventral striatum does state-value learning, this broader framework implicates a broader set of areas including the

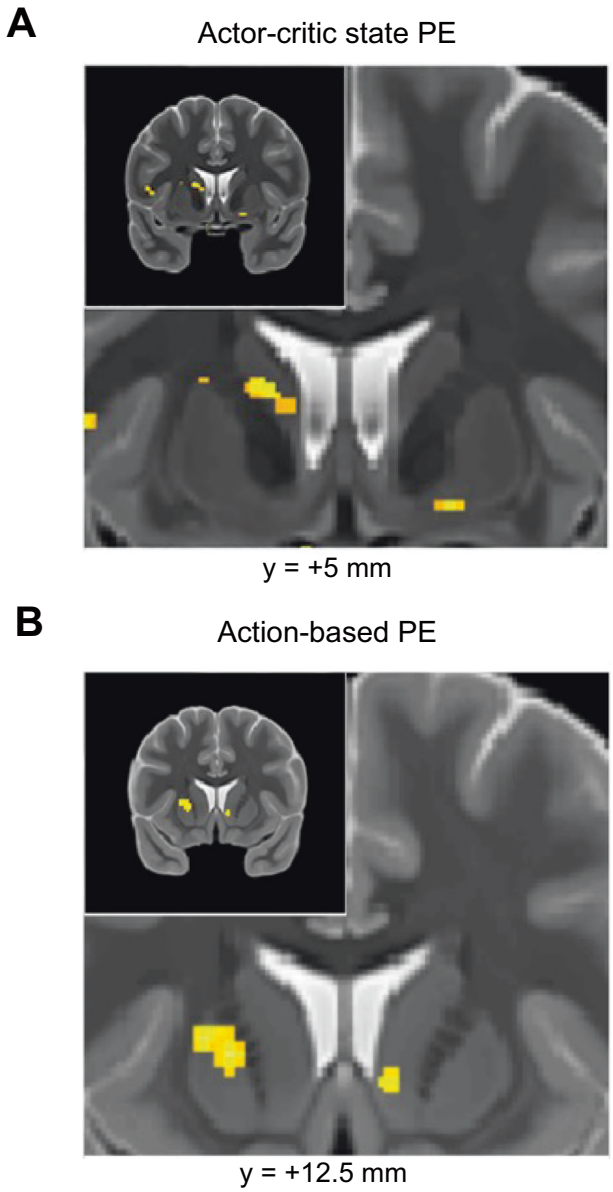

Fig. 2 Evidence for both actor-critic state prediction errors and action-based prediction errors in the human brain. A Regions of both ventral and dorsal striatum correlating with the state-based prediction error generated by an actor-critic model. **B** Regions of the ventral and dorsal striatum correlating with the action-based prediction error generated by a Q-learning model. Reward prediction errors generated by both of these strategies appear to be present in the brain simultaneously in a task designed to allow these two signals to be separately measured, consistent with behavioral findings that a mix of these two strategies appear to best explain human behavioral performance on the same task. Adapted from [31].

amygdala, vmPFC/caudal OFC, ventral striatum, ventral-pallidum, basal nucleus of the stria terminalis, as well as deeper areas including the hypothalamus. This suggestion is based on the observation that this network of areas has strong interconnections with both ventral temporal areas important for object recognition and the hypothalamus which motivates behavior to satisfy physiological drives [52]. Therefore, areas across the ventral system are an interface between information about objects in the external environment, and the internal physiological state of the animal. Learning the values of objects, which is the information about what objects can do to or for the agent, therefore, likely occurs in this system. The ventral system correspondingly identifies current behavioral goals [45], because behavioral goals are driven by current physiological needs and objects in the environment that can satisfy those needs [51].

The dorsal system, composed of parietal-dorsal frontal systems and the dorsal striatum on the other hand, has information about metric spatial aspects of the environment, including object locations [53, 54]. This system can, therefore, calculate the moment-to-moment actions necessary to obtain behavioral goals. One important prediction of this framework is that action values, as they are often explored in laboratory settings, are not normally learned by the dorsal system. For example, monkeys are often trained to make a saccade in a particular direction to obtain a reward [19, 55]. Although animals can learn these tasks, and the oculomotor component of the dorsal system is engaged during this form of learning, these tasks are not ethologically reasonable. The dorsal system may play a more important role in hierarchical RL [56]. This has not been extensively explored, but sequence learning is a form of hierarchical learning, and the dorsal system does play an important role in sequence learning [57]. Therefore, this hypothesis suggests that the ventral system is important for learning state values, which define the values of objects in the environment. These state values define behavioral goals, where a behavioral goal is a future state that provides an immediate reward. This model further suggests that the dorsal system is important for hierarchical RL (see section 2 below), which is learning how to organize actions into sets that can be efficiently deployed to obtain behavioral goals.

2) MODEL COMPLEXITY DURING LEARNING

In biological systems, RL is supported by the interaction between several computational processes that vary in complexity. Model complexity in RL is based on the amount of information about the environment an agent has available, or can use, during learning. Increased model complexity entails increased algorithmic complexity. Model-free (MF) and model-based (MB) RL are two broad classes of RL models which can differ with respect to their model-complexity. Within MF RL there are both stateless and state-based models. MB RL is always state dependent. In addition, in MB RL, one may or may not have accurate knowledge of state-transition dynamics. State transition dynamics determine the states to which one transitions, when taking a particular action in a particular state. States in RL define the information that an agent needs to make a decision. If an agent is navigating in a spatially defined task, the state may be its position in space, or if an agent is foraging, the state might be the availability of food in the current patch. The state can also relate to an agent's internal state, which for a biological organism could be the level of satiation or thirst. Anything in the environment (internal or external) that an agent needs to make a decision can be part of the state.

The simplest MF RL algorithms are stateless models. Stateless MF RL models can be used to model n-armed bandit tasks, which are frequently used to study RL [58, 59]. When these algorithms are used to model action selection in bandit tasks, they maintain a running average of the value of each action, which is the reward expected for taking the corresponding action. In stateless RL the value of an action does not depend on any other aspects of the environment.

Contextual or state-dependent MF RL is more complex than stateless RL. In these algorithms, the value of a given action, and even the actions available, depends on the current state of the world. For example, when navigating around a building, an agent may not be able to move in all directions from all positions, due to walls. State-dependent RL requires one to consider not only how a given choice will return immediate reward, but also the value of the next state, to which one transitions after an action. State-

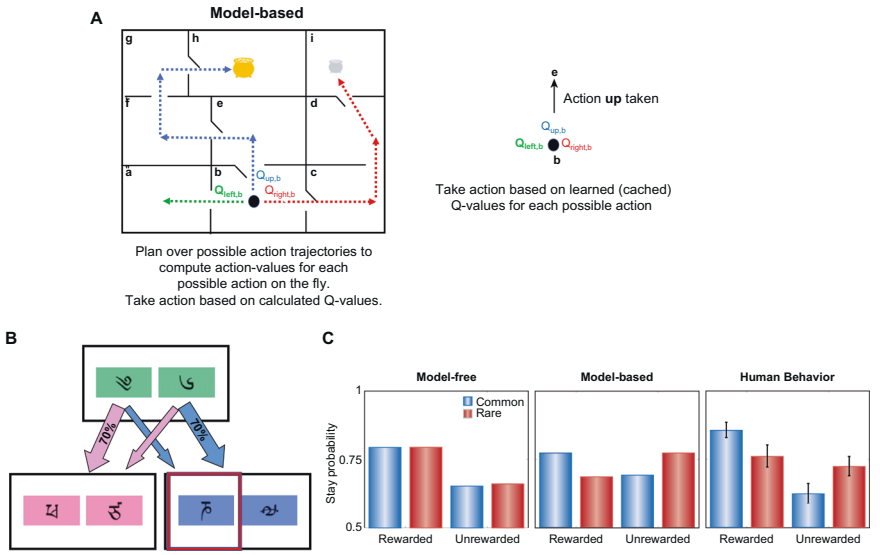

Fig. 3 Distinction between model-based and model-free RL. A Illustration of the difference between the two RL strategies. In MB RL (left) the agent utilizes an internal model of the decision problem, including the states, actions, transition probabilities between states contingent on actions and the rewards available in each state. Action-values are then computed by planning ahead and explicitly calculating the expected value of different actions based on the probability of reaching certain states and outcomes in the future. MF RL (right) uses feedback (based on reward prediction errors) to update expectations for the value of each possible action. In both models, actions with higher expected value come to be favored over those with lower expected value, leaving aside considerations about balancing exploration with exploitation discussed in the text. **B** Illustration of the "two-step" task used to differentiate model-based and model-free RL strategies in humans, and increasingly in other animals too. After taking one action (choosing either of the green stimuli), the agent then transitions to one of two subsequent states denoted by either pink or blue pairs. A subsequent choice between one of the two available pairs then results in either a reward or non-reward being received (probabilistically). Rewards received in a state after a rare transition (states arrived at with 30% probability) have different effects on subsequent behavior relative to rewards received after a common (70% probability) transition. On the next trial after a rare transition, an MB agent should be less likely to take the action leading to the rare transition, instead favoring the action leading to the highest probability of reaching that same state again, whereas a MF agent will instead be more likely to repeat the recently rewarded action even though it only rarely leads to the state that was rewarded. **C** Distinct choice patterns for MF (left) and MB (middle) algorithms are shown alongside actual human behavior (right). Choice patterns are consistent with a blend of MB and MF strategies as opposed to either MB or MF alone. Panels (**B**) and (**C**) are reproduced from [71] with permission.

dependent RL, therefore, is a much richer framework and it can model a larger class of learning problems. These models learn from experience, but they do not have knowledge of state-transition dynamics. Because these models do not have knowledge of state-transition dynamics, they cannot do forward planning and cannot predict future states.

MB RL requires an agent to have knowledge of state transition dynamics. In other words, an agent must know how its actions lead to changes in the state of the world [60]. In MF RL an agent only knows how its action in the current state will return rewards now, and in the future, but not how its actions will lead to changes in states. MB RL requires that the agent knows its current state and the future states to which they will transition for a given action. While the transition probabilities between states contingent on actions are not always known, they can be learned by agents for many real-world problems of interest including investing in the stock market and playing golf. The state transition dynamics can be captured by the probability distribution over the next state, given the current state and action.

When state transition dynamics are known to an agent, the agent can forward simulate through a state space to estimate likely outcomes for a given sequence of actions. The distinction between MB vs MF RL (Fig. 3A) was first introduced to neuroscience in order to account for the well documented dichotomy between goal-directed and habitual behavior in instrumental conditioning [61, 62]. To distinguish between these two forms of learning, animals are trained to press a lever to obtain a reward under different learning schedules (for example random interval versus random ratio). Following training, the reward is devalued, by for instance feeding the animal to satiation [62–64]. If the learning strategy has led to the development of a habit, the animal will continue to respond robustly to the now devalued action, whereas if goal-directed behavior is present, the animal will respond only minimally. When brought into an RL framework, animals that show goal-directed behavior are using a model of state-transition dynamics to simulate the outcome of their actions. Because the reward has been devalued, when the animal simulates the outcome, the animals are not motivated to press the lever. However, if an animal shows habitual behavior, they do not have a model of state transition dynamics. They have simply learned that pressing the lever is a high-valued action, and therefore they continue to press. Outcome devaluation has been utilized successfully in rodents [65, 66], monkeys [67] and also in humans [68, 69] as a means of delineating corticostriatal circuits involved in goal-directed and habitual control. This work has shown evidence for dissociable dorsal striatal sub-regions for goal-directed and habitual learning respectively (reviewed extensively elsewhere e.g., [11, 70]).

Another assay of MB vs MF processing is the so called "two-step task" [71] (Fig. 3B, C). With this task, it has been found that human behavior appears to be a mix of both MB and MF strategies.

In humans, encoding of prediction errors in overlapping regions of the anterior ventral striatum were found to exhibit a mix of MB and MF information [71], suggesting that part of the ventral striatum is involved in both MB and MF RL. Subsequent studies have implicated the human equivalent of the rodent dorsolateral striatum, the posterior putamen and the adjacent globus pallidus, in encoding MF representations in the form of prediction errors and/or value signals [72–75].

One major challenge in the study of MB and MF RL is that following repeated performance of a task such as the two-step, it becomes possible to express behavior that looks entirely model-

based even though it is model-free, so long as the state-space representation is complex enough to encode dependencies between trials [76]. Conversely, an impoverished or incorrect model-based strategy can end up looking model-free [77]. In practice, human participants or animals can potentially pursue several different strategies which vary in their assumptions about the structure of the state-space, some of which are MB and some of which are MF, and in many instances, it can be challenging to identify which are which on the basis of typical behavioral assays.

An exciting (though speculative) possibility is that there may be multiple behavioral strategies in play even within the same individual, and even simultaneously, as opposed to the simplified distinction between a single MB and a single MF strategy. Thus "model-based" and "model-free" behavior may describe a family of strategies that exist along a continuum [78, 79]. We expand on this point later in the context of a discussion about arbitration between strategies.

Though the MB and MF distinction is typically studied in an instrumental context, the dichotomy may be applicable to Pavlovian behavior too [13, 80]. Pauli et al. [81] had human participants perform a sequential Pavlovian learning task in which sequences of stimuli were paired with either a juice reward or a neutral outcome. Using multivariate analyses, it was possible to identify the contribution of sub-regions of the striatum to different forms of knowledge about an outcome i.e., whether the stimulus has value by virtue of being associated with a reward, or via knowledge about the identity of the outcome, i.e., which stimulus was associated with juice or a neutral outcome. Decoding accuracy in the ventral striatum was correlated with value-based changes in behavior consistent with either MB or MF generated value signals, while decoding accuracy in the dorsal anterior caudate was correlated with knowledge of the predicted outcome (whether a reward or a non-reward). This latter knowledge may be an exclusively MB process because it requires representation of a cognitive map of the associations. This raises the possibility that learning processes can be categorized along several dimensions. One axis is the target of the learning (e.g., state vs action) as discussed earlier, and the other is the algorithmic operation that is used to implement the learning, which is considered in this section. Thus, the distinction between MB and MF RL may go beyond the classic dichotomy between two forms of instrumental conditioning to which the computational theory was first applied. However, some forms of Pavlovian behavior may not be well explained by either MB or MF RL dichotomy, such as when some Pavlovian behaviors based on sensory features of an outcome are devaluation insensitive [82], suggesting that even a two-dimensional scheme may not be sufficient to capture the diversity of learning algorithms in the brain.

Although knowledge of state-transition dynamics brings one minimally into MB RL, such knowledge by itself does not solve the full learning problem. Even if one knows the state transition dynamics and the immediate reward pay-out of each state, the values of each state and the actions available in those states must be determined, using for example policy iteration or value iteration [2], which are algorithms for learning in MB RL. Therefore, in full MB control, state values, action values, and state transition dynamics are known. These models can be used to characterize theoretical performance on simple bandit tasks, and in this case, they allow one to optimally manage the explore-exploit trade-off [23, 83].

Within MB RL there are also Markov Decision Processe (MDP) models and Partially Observeable Markov Decision Process (POMPD) models. In MDPs the state is observeable and this information can be directly used to select an action. In POMDPs the states are not directly observeable and must be inferred. With POMDPs, the state is inferred from data which is observed and one therefore works with a probability distribution over states. A common example of a POMDP process is the standard binomial bandit task often used in RL experiments. In a binomial bandit, the underlying reward probability of an option is not known. It must be inferred from the reward outcomes received when the option has been chosen. One could also conceptualize this as being presented with a bandit chosen from a set of possible bandits. After choosing each option and receiving outcomes, one would be able to build a probability distribution over bandits, and use this distribution to estimate the necessary values.

HIERARCHICAL RL

An inherent challenge for RL in both biological and non-biological systems is how to deal with the enormous richness of both the state-space and the multiplicity of actions available for selection in that space. Consider the actions involved in moving from your sofa to your refrigerator to obtain a soda. An RL agent would need to learn about and perform a long sequence of actions to accomplish this task. But at what level of granularity should an action be considered and learnt about? Should an action consist of a specific sequence of muscle contractions? Should it be considered more abstractly at the level of moving your left leg, or more abstractly again as walking one step, or even at the level of getting up off your sofa, and walking to the door? Clearly actions can be described at many different levels of abstraction, and the nervous system must deal with all of them. Algorithmically, however, If actions are considered at too high a granularity (such as the level of muscle contractions) then the RL agent will suffer from the curse of dimensionality in which the space of possible actions is so vast, it cannot efficiently learn about all the possible actions in a reasonable time frame, and/or tractably plan over those actions. Hierarchical RL (HRL) is one proposed solution for this problem [84].

To achieve this abstraction in HRL, elementary actions can be clustered together into "options", which are sequences of actions. The actions that compose options are learned about by having sub-goals that can generate pseudo-rewards if obtained. So, for instance, the sub-goal of getting up off the sofa generates a pseudo-reward, as would getting to the kitchen, and opening the refrigerator etc. Each of these sub-tasks become defined as an option, and the value of these options toward obtaining the overall goal of the agent are then learned about as if they are elementary actions in standard RL. Thus, in HRL, learning happens on multiple levels. At the top of the hierarchy is the overall goal of the agent, for instance to obtain a particular reward such as a soda. Lower levels of the hierarchy learn the more fine-grained sequences of actions that would need to be pursued in the service of this overall goal. The framework of HRL has considerable appeal when it comes to understanding how the brain solves complex tasks in which it is necessary to bridge from the performance of granular motor actions such as a particular muscle contraction, up to high-level action sequences such as "get up off the sofa". However, our understanding of the neural mechanisms for implementing hierarchical RL is still in its infancy. Some evidence has emerged for the neural coding of pseudo-rewards in the striatum [85], consistent with a HRL framework. Theories of prefrontal function fit naturally within the notion of hierarchical abstraction instantiated in HRL [86–88].

Although there has not been a lot of work on hierarchical learning in behavior, there is a long history of work on action sequences [89]. Sequential behaviors are an important hierarchical mechanism for organizing actions. Neural correlates of sequence onset and termination, likely related to chunking (i.e., the process of dividing long sequences of actions into chunks- perhaps corresponding to options), has been found in dorsal-lateral prefrontal cortex (dlPFC; [90]) and the striatum [91]. In the dlPFC work, it was shown that neurons had a specific phasic response before and after the execution of a sequence of actions, in contrast to frontal-eye-field neurons that responded to the

individual movements in the sequence. Similarly, in [91], the authors found that primary motor cortex represented individual movements from a sequence, but the striatum responded specifically at the beginning and end of a sequence of actions that had been reinforced. When unreinforced movements that formed parts of the sequence were carried out, the striatum also did not show activity before and after the sequences. Other work has shown that complete sequences of actions are planned in parallel before execution in dlPFC [92, 93]. Thus, dlPFC and the dorsal striatum appear to delineate reinforced sequences actions, and this activity may underlie sequence chunking, a form of option discovery.

Formidable challenges remain in understanding how these HRL mechanisms might be implemented both computationally and within cortico-striatal circuits. For instance, what computational principles does the brain use to segment a task problem into options and sub-goals [94]? One idea for this, is that the brain exploits natural breakpoints or information bottlenecks [95]. How many different levels of hierarchical organization and/or abstraction are appropriate? How does the brain generate separate prediction errors for implementing learning at each level of hierarchy? Another important question concerns how HRL interfaces with MB and MF RL. Are different levels of the hierarchy differentially sensitive to MB and MF control? For instance, one proposal is that the top level of the hierarchy is MB while lower levels are predominantly MF [96]. We speculate that MB and MF control might operate at many different levels of the hierarchy, depending on the task at hand, and depending on the reliability of those strategies, as discussed further below.

3) DISTINCT RL PROCESSES FROM PREDICTION TO DECISION-MAKING AND ACTION-SELECTION

Once values for actions or stimuli have been learned, these signals need to be leveraged for the purpose of deciding which action to pursue. We now consider the brain mechanisms of this action-selection process.

Selecting a specific action out of a range of alternatives depends on the ability to compare and contrast the values of actions, so that all else being equal (leaving aside the exploration/exploitation tradeoff discussed below) the action with the highest expected value is chosen. The process of selection between alternative actions has been widely studied in decision-making. This animal literature has focused mostly on decision-making about perceptual attributes of a stimulus, such as the prevailing direction of motion of a series of moving dots [97], although some studies have focused on the neural mechanisms of decisions between stimuli or actions with varying reward value [48].

Much of the decision-making literature has emphasized the role of the cortex in decision-making [97–99] including the posterior parietal cortex [97] and prefrontal cortex [100]. The dominant proposal about how decisions are made in the brain is via evidence accumulation as featured in computational models of decision making such as the drift diffusion model [101]. According to such models, noisy evidence is accumulated for particular decision options by different pools of neurons until a bound or threshold is reached by which the decision-maker opts for one or other option. Consistent with this theoretical framework, studies in monkeys and rodents have found that neurons in lateral intraparietal sulcus and prefrontal cortex exhibit ramping activity, which is presumed to correspond to the evidence accumulation delineated in such models [100, 102, 103]. Though these models have been applied successfully to perceptual decision-making in which multiple discrete samples from a noisy percept are used to update evidence, the same processes have also been adapted to value-based decision-making where it is assumed that the subjective value of stimuli or actions are repeatedly sampled [104]. Evidence for accumulation in parietal and prefrontal regions has also been found in humans, using both fMRI and surface EEG in both perceptual and value-based decision-making [105–108].

While much of the decision-making literature has focused on cortex, especially in perceptual decision-making, another parallel literature has emphasized the contribution of the striatum in action-selection and performance. According to some classic theories of neural RL, the cortex represents the available actions, the striatum represents the values of those actions, and mid-brain dopamine neurons signal reward prediction errors, which update action values in the striatum. Action selection then takes place either within the striatum, or in basal ganglia output [109] structures including the globus pallidus [110, 111]. The selected actions are then fed back to cortex, via the thalamus, and the cortex executes the selected action via its descending projections. From the evidence reviewed so far, this theory is likely too simplistic.

Traditional views on dorsal striatum contributions to action selection have focused on the so-called direct and indirect pathways, associated with striatal medium spiny neurons (MSNs) expressing D1 or D2 dopamine receptors, respectively. Direct MSNs have been implicated in implementing motor actions, and indirect MSNs in inhibiting those actions [112, 113]. Recent studies using molecular tools to selectively label MSN neurons suggest that those classical models of striatal action-selection need to be expanded and revised [114]. Both direct and indirect neurons appear to similarly encode actions during the performance of motor activity [115, 116], rather than being active or inhibited during movement and rest, respectively. Instead, these distinct neuronal populations may correlate with action-values differently, with direct neurons correlating positively and indirect pathway neurons correlating negatively with action-values [117–120]. These experimental findings fit with a theoretical proposal that direct pathway (D1) MSNs code for positive action values and indirect pathway (D2) MSNs code for negative action values, respectively [121], suggesting that these neurons play specific roles in decision-making about whether to take an action or not as a precursor to the implementation of actions. Congruent with this notion is the finding that dorsal striatal neurons exhibit ramping behavior during decision-making consistent with evidence accumulation, similar to that found in the cortical areas reviewed previously [100, 122]. If the dorsal striatum is involved in perceptual decision making, however, it is not via a dopamine dependent process, as injections of dopamine antagonists into the caudate have no effects on perceptual inference, but they do affect choices driven by learned values [29].

The dorsal striatum may also be differentially engaged as a function of whether choice is being guided by a MF RL strategy or not. Jessup and O'Doherty [123] compared performance of human participants on a gambling task in which behavior was either consistent with MF RL (i.e., choosing a stimulus that was previously rewarded), or with a different strategy known as the gambler's fallacy whereby participants assume that the more they choose from a particular action the less they will be rewarded. On trials in which participants chose consistent with MF RL, activity was greater in the dorsal striatum relative to when participants chose consistent with the gambler's fallacy. Moreover, this region of dorsal striatum overlapped an area correlating with MF RL prediction errors, suggesting that the same striatal region involved in using prediction errors for learning is involved in deploying that learned strategy to guide behavior on a trial by trial basis [123].

There is, however, an on-going debate about the relative importance of the basal ganglia vs. the cortex in action selection. Theoretical models, and some experiments in rodents have suggested that the basal ganglia may carry out action selection [124, 125]. In many cases that have been examined, although rarely in the context of learning, cortex represents chosen actions before the striatum [27]. Furthermore, neurons in the globus pallidus (a basal ganglia output structure), appear to represent an

urgency signal during decision making, and not the selected action [126]. Even if the basal ganglia is involved in action selection, which might be the case in some oculomotor tasks, it is the descending projections to the substantia nigra pars reticulata, which project to the tectum (i.e., superior colliculus) that likely mediates action selection, not return projections to the cortex [127]. Similar results apply for other motor behaviors, in that descending projections from the striatum to the mid-brain reticular formation are important for action selection [128]. Moreover, it is also not clear where values are stored and updated. Although much work suggests that the striatum is important for this process, other work suggests that plasticity during learning may be wide-spread. Another issue is that most decision-making studies in animal models have not examined decision making in the context of learning, where a decision must be made on the basis of past outcomes. Rather, the decision making literature typically focuses on choices when all information is present on the screen, or perhaps has been presented within a few seconds. Therefore, whether these decision mechanisms generalize to decisions over learned values is not yet clear.

When taking this literature as a whole, it is currently unclear where in the corticostriatal network decisions are made. One possibility broadly consistent with the experimental evidence to date is that action-selection arises at the network level through multiple interacting brain regions, and specifically cortico-cortical and cortico-striatal interactions. Considerable additional work will be necessary to understand how the processes involved in RL are distributed across cortical-basal ganglia-thalamocortical networks.

Action selection and the explore-exploit trade-off

In MF RL, agents must manage the explore-exploit trade-off. For example, when you move to a new city, you need to try several restaurants to decide which you like. After finding a few restaurants whose food you prefer, you have to decide whether you want to continue exploring new restaurants, or return to exploit your previous favorites. This interplay between exploring unfamiliar options and exploiting options of known value is fundamental to optimizing reward in the future. Exploration always requires foregoing an immediate reward, to learn about an option that may be better than familiar options. Therefore, exploration is a trade-off between immediate and future expected rewards [83, 129].

When full MB RL is applied to bandit tasks, it specifies an optimal solution to the explore-exploit problem. Because these models are optimal, under some assumptions, they provide a ground truth to which behavior can be referenced. These models also decompose state and action values into the immediate and future expected value of an option. The immediate expected value is the reward that is expected immediately when an action is taken in a given state. The future expected value is the reward that is expected, over some relevant future time horizon, when the optimal policy is followed in the future. These two components, derived from the model, can be regressed on behavior and neural activity to see the extent to which each affects choices and neural responses [23, 24]. Theoretical work has shown that exploration is driven by future expected rewards, because these are rewards one can expect to obtain in the future, following exploration. Exploration is also driven by uncertainty. When the values of all options are known, there is nothing to explore. Exploration is only valuable when there are available options with unknown reward distributions. Unknown reward distributions associated with choice options can arise for many reasons including environments in which choices have non-stationary reward distributions or new options are periodically presented. Behavioral work has shown that exploration can be driven by both direct and undirected exploration, where undirected exploration is non-specific noise in the decision process, and directed exploration selects options about which there is more uncertainty [58, 87, 129].

The explore-exploit trade-off has been studied using several paradigms. Within these paradigms, two approaches have been used to drive uncertainty, and therefore exploration. The first study to examine the explore-exploit trade-off used drifting bandits [58] to drive uncertainty. In drifting bandit paradigms, the rewards associated with choices are non-stationary. Because the rewards are non-stationary, the uncertainty associated with an option increases when it is not sampled. The only way to maintain an estimate of the reward distribution associated with an option is to sample it. This study used a MF approach to identify when subjects were exploiting options of known value, vs. exploring options that had not been sampled recently. They found that exploration activated the intra-parietal sulcus and the frontal pole, whereas exploitation engaged the vmPFC. Additional studies, using a bandit paradigm in which the amount of uncertainty about options was systematically controlled, found that TMS applied to frontopolar cortex disrupted exploration [130]. Experiments based on drifting bandit paradigms in monkeys [131] have found that frontal-eye-field neurons predicted spatial choices when animals were exploiting known options. However, when animals switched to exploring options, FEF neurons no longer predicted choices.

Exploration has also been studied using paradigms in which familiar choice options are occasionally replaced with novel choice options ([23, 132], Fig. 4). When this occurs, subjects have to decide whether to explore the novel choice option, or continue to select the remaining familiar options of known value. The novel options have unknown reward distributions, and therefore they must be sampled to learn whether they are better than the other options. Recordings in orbitofrontal cortex (OFC), the ventral striatum and the amygdala found that all three areas coded the values of exploiting known options, but also the value of exploring a novel option ([23, 24] Fig. 4C). However, the amygdala tended to encode exploration value more robustly than the other areas, and may be playing an important role in driving exploration. Whether the frontal pole, which played a role in exploration in human studies, would also be involved in exploration driven by novelty is not clear. Overall, the neural circuitry underlying the explore-exploit trade-off is not well understood. Whether different paradigms, which drive exploration through different mechanisms (e.g., novelty vs. non-stationarity), engage the same circuits is not well understood. Furthermore, the role of different cortical and sub-cortical circuits in different aspects of the explore-exploit problem is also not well understood. Therefore, additional work will be required to clarify these issues.

4) STATE-SPACE REPRESENTATIONS

RL models whether MB or MF (except state-less RL, see above), require internal representations of the possible states of the world that an agent visits, and of the actions available in those states. This information is the scaffold upon which action-values and state-values are learned and utilized at the time of decision-making. Identifying appropriate and useful state-space representations is a computationally challenging problem that has been a major barrier to progress in the application of RL to artificial intelligence. Arguably, understanding how the brain accomplishes this is also one of the hardest problems in biological RL. To gain an appreciation of the immensity of this challenge consider the extremely high dimensional nature of the sensory inputs received by an individual walking down a busy city street. Somehow, the brain must rapidly process this vast data stream, find relevant low dimensional representations of the state of the world, and identify available actions—can you go left, right, straight ahead, press a button to cross the road, or just walk out in front of traffic? If the dimensionality of the state-space representation is too high then "the curse of dimensionality" applies. In this case learning about the value of each separate state via MF learning will rapidly

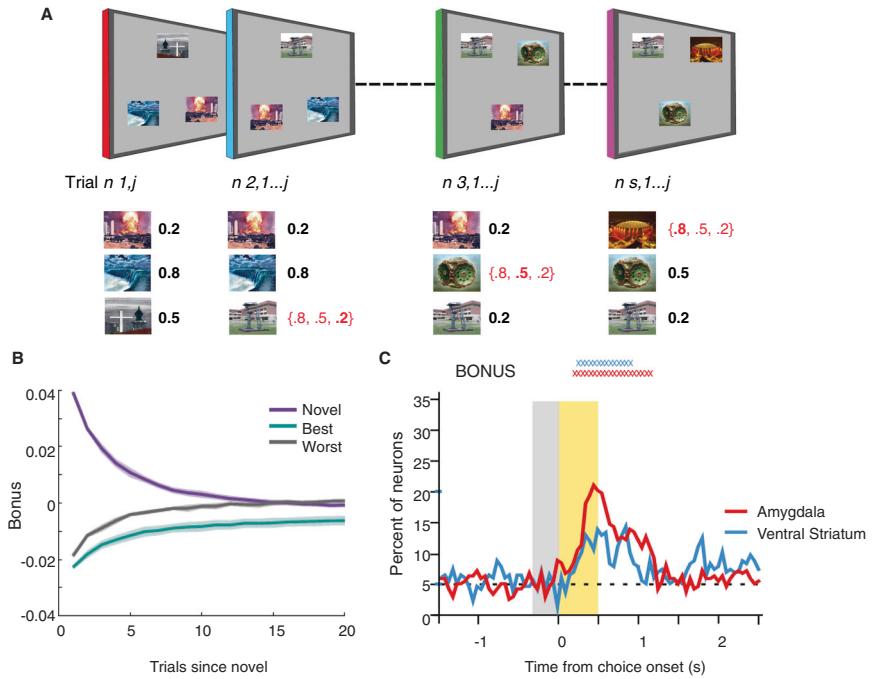

Fig. 4 Neural correlates of exploring novel options. A In the novelty task, reward probabilities were associated with visual stimuli. Periodically (every 10–30 trials) a visual stimulus was replaced with a novel stimulus. Novel stimuli had unknown reward values. The monkeys had to explore the novel options to learn their reward values. **B** Average novelty bonus values, derived from a POMDP model. Novel options had large novelty bonuses in the first few trials when they were introduced. The novelty bonus is related to how well the reward outcome of the stimulus is known, which relates to how many times it has been chosen. Best and worst refer to the best and worst familiar option, on the basis of previous reward feedback. The best options on average have the lowest novelty bonuses because they have been chosen the most. **C** Neural activity in the amygdala correlates with the novelty bonus of the chosen option more than activity in the ventral striatum.

become intractable, as will planning efficiently in a massive MB state-space. At the same time the state-space has to be of sufficient granularity to facilitate learning of useful policies. The state-space identification problem can be thought of as the fundamental problem faced by our sensory systems—we need to identify the states of the world and the actions available in the world so that RL can operate on those representations. In essence this becomes the interface between brain systems involved in perception and those involved in value-based decision-making.

To date, several brain systems have received focus for their role in state-space representation. The hippocampus is perhaps the most studied especially for its role in encoding spatial information that can be used for the purpose of guiding RL. Place cells in the hippocampus can provide a veridical representation of where an animal is in a given environment, which can then be used as an index of the current state and facilitate planning about future states [133, 134]. The hippocampus has been found to be especially involved in MB RL [135], perhaps because it is capable of supporting representation of the model (or cognitive map) underpinning MB learning [136], but it also may play an active role in MB planning [137, 138]. It is also possible that hippocampal place cells may play a role in encoding expectations about future state-occupancy—which would enable outcome-value sensitive action selection without necessitating full MB planning nor requiring encoding of a fully elaborated state-space [139–141].

Beyond spatial coding, other information about states could include the identity of specific objects or people. A state space is most generally a graph, where states are nodes and edges join nodes between which one can transition, depending upon chosen actions and the environment. The hippocampal complex may also contribute to the representation of non-spatial cognitive maps [142] or even social-networks [143]. While the hippocampus is likely to be an important component of the brain's state-space apparatus, other brain areas may also play an important role.

The OFC has also been focused on for its potential contribution to state-space encoding. In particular the OFC is known to encode representations not only of value and reward, but also the identity of stimuli and outcomes [144, 145]. Such associations may form part of the basis by which the value of potential outcomes or goals can be retrieved during MB planning. The OFC has also been implicated in inference over hidden states. While in some situations, the current state of the world may be directly observable (i.e., I am in the kitchen vs. in the living room), in complex situations it may not be immediately obvious what state one is in, and instead the brain needs to infer or predict the state. Wilson et al. [60] argue that the OFC plays a special role in signaling what state of the world an agent is in when states are not directly observable but instead must be inferred.

The proposed contribution of OFC to state-space encoding is sometimes viewed as being in contradistinction to the proposal that the OFC is involved in encoding value and reward. There is a very strong literature implicating neurons in the OFC in the encoding of value signals [46, 146] and in encoding features from which an overall expected value signal can be constructed such as the magnitude or probability of an outcome [147] or other attributes of a potential outcome such as its fat or carbohydrate content [148]. Our view is that the role of OFC in encoding state-space representations is not inconsistent with a role for this region in encoding value and its precursors. Rather, we suggest that the role of the OFC in encoding and performing inferences over state-spaces, is in the service of using information about the state of the world, or the context an agent is in, alongside features and attributes of potential outcomes (including stimulus-stimulus associations), in order to compute an overall context-specific value signal for a particular outcome or goal, which in RL terminology would correspond to the reward signal [149].

Reversal learning has long been used to study response inhibition and behavioral flexibility [150, 151], and is thought, at

least in some species, to be an important function of OFC [152]. Reversal learning can be viewed as a state inference process. In reversal learning paradigms, reward is first associated with one of two options. After subjects have learned to select that option, the choice-outcome mappings are switched, and subjects have to switch their choice preference. This switch in choice preference could be mediated by an incremental MF learning mechanism. However, reversal learning can also be mediated by a state-inference process [60, 153, 154], where the currently rewarded option defines the current hidden state of the world. In recent work [155], examined behavioral and neural correlates of reversal learning in over-trained primates. In this experiment reversals occurred within a predictable window of trials. Earlier work found that monkeys could develop a prior estimate of the reversal interval, and this prior could be combined with the trial-by-trial evidence for a reversal, to drive choices [156]. In the recent study, Bartolo and Averbeck found that monkeys rapidly switched their choice preferences, when they detected reversals. The choice behavior of the monkeys, when they reversed, was better modeled using a Bayesian switching model, than a MF learning mechanism (see also [153] for similar evidence in humans). While the animals carried out the task, they also recorded the activity of 500–1000 dlPFC neurons, simultaneously. Examination of the population activity showed that there was a clear signal in the dlPFC on the trial in which the animals switched to the new option. Furthermore, this signal developed following the (usually) unrewarded choice-outcome in the previous trial. The dlPFC signal could also be used to predict accurately the trial on which the monkey switched its choice preference. Therefore, there was a clear dlPFC correlate of the behavioral state-switch shown by the animals in the reversal learning task.

The role of rodent vmPFC (i.e., infralimbic/prelimbic) has also been explored in hidden state-inference [157]. In these experiments, rodents were trained on a task in which an odor cue predicted reward delivery at an uncertain time. The task was structured such that the animals could infer that they were either waiting for a reward within a trial (the certain reward condition), or may have transitioned to the inter-trial interval (the uncertain reward condition). The authors examined the activity of dopamine neurons, and found that reward prediction errors (RPEs), and correspondingly the activity of dopamine neurons, decreased with time when the animals were waiting for a reward in the certain condition. However, RPEs increased with time when the animals thought they had transitioned to the ITI in the uncertain condition. They then examined the effects of inactivating mPFC, on the responses of dopamine neurons. They found that following inactivation, responses of dopamine neurons were unchanged in the certain condition. However, responses no longer increased with time in the uncertain condition, which suggests that inference over the hidden state, i.e., whether the task had transitioned to the ITI, was disrupted. In humans, activity in the vmPFC was found to correlate better with a state-based inference model than a MF RL agent during a reversal learning paradigm [153].

Finally, another important set of structures implicated in state-space representation is the posterior parietal cortex. Glascher et al. [158] found evidence for the presence of learning signals in these structures as well as in dlPFC that could underpin the acquisition of state, action, and state transitions (Fig. 5a). These are key components of the state-space needed for MB inference. Further building on this, in a recent study [159] had humans play Atari games while being scanned with fMRI. These authors utilized deep RL models in which deep-convolutional neural networks are married to RL to solve the state-space identification problem in artificial intelligence applications. This model was also trained to perform the same Atari games that the humans played. The authors examined the representation of the relevant game states in the network layers, as well as in the brain. They found similarities between how the deep network represented relevant states of the game, and how the brain did. In particular, a notable aspect of the parietal cortex representation was that it became invariant to changing features of the game environment that were not relevant to game play, unlike the early visual cortex which was sensitive to changes in irrelevant sensory features. Thus, the posterior parietal cortex appears to be involved in abstracting the relevant sensory features necessary for building an abstract state-space representation needed for RL.

Taken together, these findings support contributions of multiple brain regions to the thorny computational problem of state-space identification, abstraction, and inference, which is also related to questions in hierarchical RL. Understanding how these brain areas work together to facilitate state-space identification and support learning over those states and actions is still an outstanding problem. It remains possible that the brain's RL system can flexibly use different forms of state-space representation depending on the problem at hand and that different parts of the brain may contribute differently depending on the nature of the problem. For instance, if a task depends heavily on spatial information, the hippocampus will be involved (even though this structure is known to contribute to non-spatial coding); or if a task depends heavily on rapidly selecting actions in a fast moving scene, parietal cortex will be engaged; or if computations about hidden states are required, OFC will participate. In practice these areas likely all contribute seamlessly to facilitate efficient state-space identification in an integrated manner. An interesting question for the future is where and how is this information integrated, and whether the striatum plays an important role in this integration process, above and beyond its role in value-based learning and selection that takes place once the relevant states and actions have been identified.

5) ARBITRATION BETWEEN DISTINCT RL MECHANISMS

The operation of distinct MB and MF mechanisms, or even the possible operation of multiple MF algorithms such as the actor-critic and direct action-learning, suggest the need for a mechanism to provide oversight over the deployment and regulation of these strategies. The notion that different strategies compete for the control of behavior has a long history in dual-system theories within psychology. Of particular relevance, the control over behavior by goal-directed and habitual systems has been found to depend on several variables including the length of training and the specific reinforcement schedule [63, 160]. Inspired by those findings, Daw et al. [61], proposed the existence of an arbitration mechanism that takes into account the degree of uncertainty in the predictions of the MB and MF systems to determine the control over behavior of these systems, so that the system with the most reliable predictions dominates. Building on this initial proposal, a number of different theories of arbitration have emerged. Lee et al. [75], proposed that uncertainty in the two systems can be cheaply approximated by keeping track of the reliability of the two systems, estimated through the recent history of prediction errors accumulated in those systems. For the MB system, this was proposed to be accomplished via keeping track of the state-prediction errors used for learning the transition model, while in the MF system, via keeping track of the average of the reward prediction error signals. Alternative theories have emphasized other potentially important considerations about the arbitration process, such as the relative gain in expected value of pursuing a particular strategy [161, 162], the value of obtaining additional information through MB means [163], the trade off between speed and accuracy [164], or the cognitive or computational cost of implementing a particular system [140, 142]. Implicit in these ideas is that there is a trade-off that must be resolved between the amount of cognitive effort expended and the gains in either expected value or accuracy from pursuing a more cognitively demanding MB strategy over a MF one.

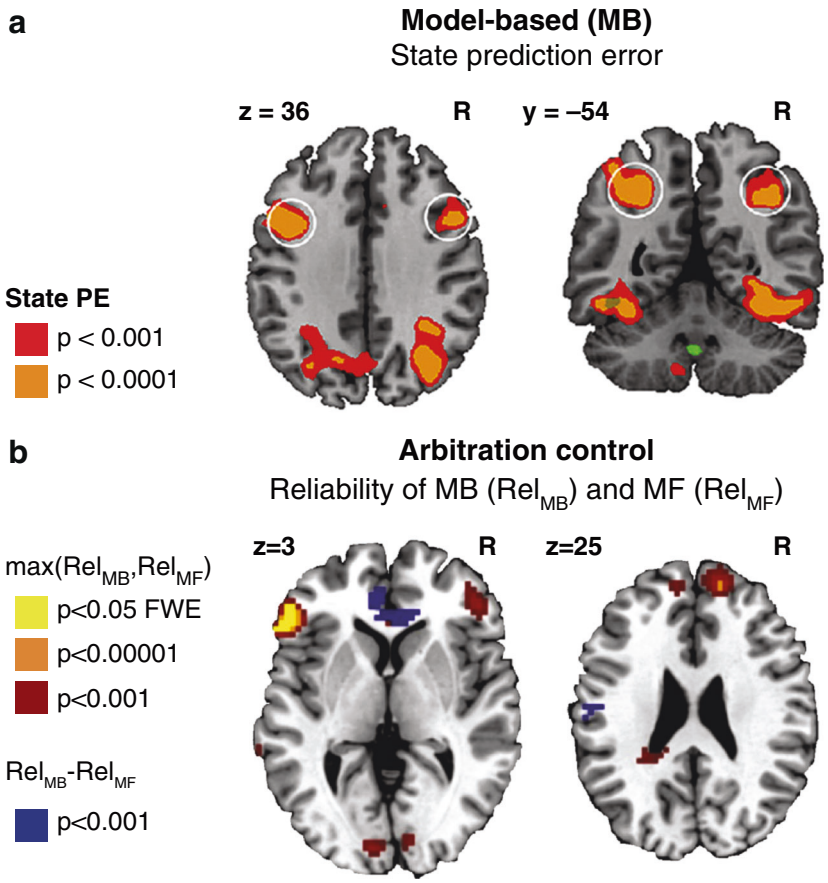

Fig. 5 Model-based learning and arbitration between MB and MF in the brain. a Evidence for the encoding of state-prediction errors in the dorsolateral prefrontal cortex (left panel, circled) and posterior parietal cortex (right panel, circled) which could underlie learning of state-action-state transition probabilities. Adapted from [158]. **b** Evidence for the role of ventrolateral prefrontal cortex in encoding the reliability of the model-based and model-free strategies, consistent with a putative role for this region as an arbitrator between these strategies. Adapted from [75].

While the jury is out about which of these arbitration mechanisms best explains behavior, it should be noted that there are a number of challenges that arbitration theories will need to accommodate. First, estimating the expected value of pursuing a particular system in order to arbitrate over that system requires actually implementing that system to calculate the expected value in the first place, thereby mitigating any potential cost savings that would be obtained from avoiding doing so. Second, estimating cognitive cost can be computationally challenging to do, although it may be possible for relatively simple heuristics to be used to approximate that (such as decision time [165]). Third, it is not the case that a more cognitively complex MB system will necessarily be the most reliable or accurate compared to a MF system. This is true only in the special case where the MB system has access to an accurate transition model of the world, and in which MB calculations are not subject to any cognitive constraints. Neither of these situations likely routinely apply in the real world. A MB system can end up providing very poor predictions if it has an inaccurate or incomplete model of the world. Given all of these caveats, one of us [79] has argued for the parsimony of a pure reliability based arbitration scheme which can implicitly accommodate many of these considerations: for instance the penalizing of cognitive complexity is addressed by the bias/variance trade-off in which a more complex model will tend to overfit the training data thereby performing poorly in out-of-sample contexts, and in which cognitive constraints will also impact on the accuracy of a cognitively expensive system thereby decreasing its reliability.

In the brain, several studies have reported reliability signals for MB and MF learning that could potentially be used to drive the arbitration process, most notably in the ventrolateral prefrontal and frontopolar cortices (Fig. 5b) [66, 67, 141]. Connectivity analyses have found that when the arbitration model predicts that behavior should be more MB, coupling is increased with areas containing MF signals in the posterior putamen, suggestive of a mechanism for dynamically downweighting the contribution of MF RL on behavior as a function of the arbitration process via a cortico-striatal route. Consistent with this proposal, several transcranial direct stimulation studies have shown that stimulation over the ventrolateral prefrontal cortex can modulate the balance between MB and MF control, such that ventrolateral prefrontal excitation produces more MB control and ventrolateral prefrontal inhibition yields more MF control [166, 167]. These results are consistent with a role for anterior prefrontal cortex in implementing the arbitration between different systems such as MB and MF RL.

It is also possible that MB and MF strategies influence each other directly—such as for instance an MB mechanism training up an MF strategy, such as where credit assignment in the MF system is guided by a model [81, 168].

O'Doherty et al. [79] has recently argued for a more general role of these areas of prefrontal cortex in regulating the operation of many different strategies or "experts", not just singular MB and MF strategies (related proposals have been put forward by Badre and Frank [169] and Doya et al. [170]). According to this idea, the brain is composed of many different experts which operate on either

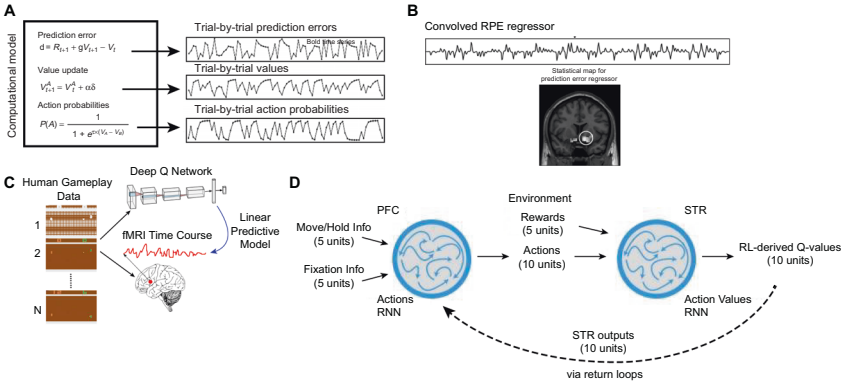

Fig. 6 From algorithmic to implementational-level modeling of neural RL mechanisms. Illustration of algorithmic level modeling of neural data (here shown for fMRI data analysis). An RL model is fit to behavioral data and parameter estimates are derived which are then used to generate model-based time series capturing time-varying effects of a particular model-based variable, e.g., reward prediction errors. **B** These signals are then convolved with a hemodynamic response function to account for hemodynamic lag and regressed against fMRI data. Adapted from [183]. **C** Another approach is to use a connectionist neural network model such as a feed forward deep-Q network, train this model on the same task as the subject is performing (the example shown here is playing the Atari game "Pong"), and then correlate (dimensionally reduced using e.g., PCA) distributed activity patterns generated in the different layers of the neural network against fMRI data (or indeed any neural data). Thus distributed activity in the networks units can be compared and contrasted against activity in the brain. Adapted from [159]. **D** Recurrent neural network models can also be used to study the ways in which distributed architectures adapt during reinforcement learning. In recent work we built a recurrent neural network model representing frontal-striatal systems that learned sequences of actions. We explored the way learning drove changes in dynamics across these circuits and found that learning drove dynamical trajectories representing action sequences further apart in low-dimensional spaces representing the activities of model neurons.

different input data using similar algorithms, or operate on the same input data with different algorithms. This multiplicity of expert systems allows the brain to poll different advice, and by weighting this advice by its reliability, it becomes possible for the brain to develop an informed behavioral policy guided by the collective wisdom of its different underlying experts. Beyond simple MB and MF RL, other example expert(s) might be expert systems involved in implementing Pavlovian reflexes [171], or different experts involved in learning from observing others [172], or indeed multiple MB and MF strategies that make different assumptions about the structure of the underlying state-space. Within this broad framework, it is possible that ventrolateral prefrontal cortex and adjacent areas of anterior prefrontal cortex play a domain general role in allocating weights over different experts in order to compute an overall behavioral policy [79]. It remains an open question precisely how this domain general arbitration process is implemented at the level of neural circuits. We speculate that one possible mechanism is for the prefrontal cortex to selectively gate experts that are deemed to be less reliable by the arbitrator, via a cortically mediated inhibition of specific striatal circuits responsible for implementing that particular expert. Recent empirical evidence has emerged of role for a mixture of experts framework in accounting for functional specialization within the striatum [173].

CHALLENGES AND FUTURE DIRECTIONS

In this review we have highlighted five key domains in which progress has been made over the past decade in elucidating the neural computations underlying RL in cortico-striatal circuits. In this section we highlight several emerging challenges that the field faces going forward.

The first of these will paradoxically become more of a problem the more success the field enjoys in delineating the basic circuits underlying RL and decision-making. The temporal and spatial resolution and causal precision available in non-human animal models has offered unprecedented opportunities to dissect neural systems and circuits to a level of detail never before attained. That increasing detail indubitably brings new understanding. Much success has been achieved through establishing similarities in mechanisms across multiple species from mice to humans in RL and decision-making [11]. However, the more we understand the neural circuits for learning and decision-making in mice or in non-human primates, the more we will have to consider differences in computational strategies and neural implementations between species. Simply put, there are millions of years of evolution between model species and humans (~80 MY for rodents and ~25 MY for macaques), and clearly those evolutionary stages have proved pivotal for the refinement of uniquely human intelligence, including the ability to efficiently learn and make decisions across long time-scales in complex state spaces. While work in animal models will continue to provide a mechanistic foundation from which to study the human brain, it is important that methodological advances continue to happen for studying the human brain so that we can overcome limitations in spatial and temporal resolution and causal inference in human studies. In this way, differences in circuits and computations across species that will undoubtedly become more evident in time, can be understood.

Relatedly, differences in behavioral training regimes and testing contexts across species might also inadvertently tap into different cognitive and computational processes. For instance, monkeys are typically overtrained on decision-making and learning tasks, often completing many tens of thousands of trials before data collection begins, while humans are instructed on a task before beginning several hundred or at most a few thousand trials of that task. The implications of these training differences may not matter for some neural processes, for example when studying the visual system, but they could be especially marked when examining processes such as the distinction between MB and MF RL whose engagement is likely to be strongly sensitive to the level of training on a particular task. Similarly, behavioral strategies deployed by mice and humans on particular tasks might not always align. Thus, it is important to consider and evaluate the potential impact of ethological, neuroanatomical, behavioral, training, and task differences on any inferences that can be drawn.

On the use of computational models to understand the computational mechanisms underlying RL in corticostriatal circuits, while it is appropriate to celebrate the enormous success of RL models as a means of gaining insight into the neural substrates of learning and decision-making, the approach to date in neural RL has relied on algorithmic level descriptions, in which

correlations between the predictions of simple RL models and spatiotemporal activity patterns of neurons or BOLD activations have been revealed [174] (Fig. 6A, B). However, the assumption of a 1:1 mapping between an abstract computational variable produced from an RL algorithm and neural activity may become strained, especially when examining RL mechanisms in more real-world or ecologically valid tasks [159], when asking questions about the nature of representations of state-spaces [28] and action-selection during RL, or when considering the sheer complexity of the neural circuitry. Using Marr's tripartite scheme as inspiration [175], arguably what is needed is to begin to gain insight into the implementation-level. Such models would involve neural networks perhaps specialized for different tasks alongside brain regions ([176] Fig. 6D), similar to previous approaches with connectionist and neural network modeling [177–179]. Those approaches had become less popular in recent years perhaps because details of the neural circuitry were lacking in order to adequately constrain the models, and because the models themselves suffered from limitations in their capacity to model cognitive phenomena. With the emergence of deep-learning [180], some of the limitations of these models have been overcome, and given advances in molecular neuroscience we are now beginning to glean new details about neural circuitry at a level never before accessible. The time is ripe to revisit the implementational level in RL and decision-making similar to initial steps being taken in related subfields [181, 182] (see Fig. 6C, D for example applications). It is important to bear in mind that the algorithmic level remains essential in order to understand in an interpretable manner how a given problem is being solved, but extending those algorithms to the implementational level will impose important constraints on those algorithmic descriptions, and better enable us to characterize how particular neural circuits actually implement those algorithms.

REFERENCES

1. Neftci EO, Averbeck BB. Reinforcement learning in artificial and biological systems. Nat Mach Intell. 2019;1:133–43.
2. Sutton RS, Barto AG. Introduction to reinforcement learning. Cambridge,MA:MIT press; 1998.
3. Schultz, W. Dopamine reward prediction error coding. Dialogues Clin Neurosci. 2016;18:23–32.
4. Nasser HM, Calu DJ, Schoenbaum G, Sharpe MJ. The dopamine prediction error: contributions to associative models of reward learning. Frontiers in Psychology. 2017;8:244.
5. Wickens JR, Horvitz JC, Costa RM, Killcross S. Dopaminergic mechanisms in actions and habits. Journal of Neuroscience. 2007;27:8181–8183.
6. Averbeck BB, Lehman J, Jacobson M, Haber SN. Estimates of projection overlap and zones of convergence within frontal-striatal circuits. J Neurosci. 2014;34:9497–505.
7. Haber SN, Kim K-S, Mailly P, Calzavara R. Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. J Neurosci. 2006;26:8368–76.
8. Alexander GE, DeLong MR, Strick PL. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci. 1986;9:357–81.
9. Barto AG. Adaptive critics and the basal ganglia. In: Models of Information Processing in the Basal Ganglia, J. C. Houk, J. Davis and D.Beiser (Eds.), Cambridge, MA: MIT Press, 1995: pp. 215–232.
10. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 1996;16:1936–47.
11. Balleine BW, O'Doherty JP. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology .2010;35:48–69.
12. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science.2004;304:452–4.
13. Dayan P, Berridge KC. Model-based and model-free pavlovian reward learning: revaluation, revision and revelation. Cogn Affect Behav Neurosci. 2014;14:473–92.
14. Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, et al. A selective role for dopamine in stimulus–reward learning. Nature .2011;469:53–7.
15. Parkinson JA, Dalley JW, Cardinal RN, Bamford A, Fehnert B, Lachenal G, et al. Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. Behavioural Brain Res. 2002;137:149–63.
16. Costa VD, Dal Monte O, Lucas DR, Murray EA, Averbeck BB. Amygdala and ventral striatum make distinct contributions to reinforcement learning. Neuron.2016;92:505–17.
17. Taswell CA, Costa VD, Murray EA, Averbeck BB. Ventral striatum's role in learning from gains and losses. Proc Natl Acad Sci. 2018;115:E12398–406.
18. Vicario-Feliciano R, Murray EA, Averbeck BB. Ventral striatum lesions do not affect reinforcement learning with deterministic outcomes on slow time scales. Behav Neurosci. 2017;131:385–91.
19. Rothenhoefer KM, Costa VD, Bartolo R, Vicario-Feliciano R, Murray EA, Averbeck BB. Effects of ventral striatum lesions on stimulus-based versus action-based reinforcement learning. J Neurosci. 2017;37:6902–14.
20. Gillis ZS, Morrison SE. Sign tracking and goal tracking are characterized by distinct patterns of nucleus accumbens activity. ENeuro. 2019;6(2): ENEURO.0414-18.2019.
21. McGinty VB, Lardeux S, Taha SA, Kim JJ, Nicola SM. Invigoration of reward seeking by cue and proximity encoding in the nucleus accumbens. Neuron.2013;78:910–22.
22. Morrison SE, McGinty VB, du Hoffmann J, Nicola SM. Limbic-motor integration by neural excitations and inhibitions in the nucleus accumbens. J Neurophysiol. 2017;118:2549–67.
23. Costa VD, Mitz AR, Averbeck BB. Subcortical substrates of explore-exploit decisions in primates. Neuron.2019;103:533–45.e5.
24. Costa VD, Averbeck BB. Primate orbitofrontal cortex codes information relevant for managing explore–exploit tradeoffs. J Neurosci. 2020;40:2553–61.
25. Lau B, Glimcher PW. Value representations in the primate striatum during matching behavior. Neuron.2008;58:451–63.
26. Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science.2005;310:1337–40.
27. Seo M, Lee E, Averbeck BB. Action selection and action value in frontal-striatal circuits. Neuron.2012;74:947–60.
28. Bartolo R, Saunders RC, Mitz AR, Averbeck BB. Dimensionality, information and learning in prefrontal cortex. PLOS Computational Biol. 2020;16:e1007514.
29. Lee E, Seo M, Monte OD, Averbeck BB. Injection of a dopamine type 2 receptor antagonist into the dorsal striatum disrupts choices driven by previous outcomes, but not perceptual inference. J Neurosci. 2015;35:6298–306.
30. Niv Y, Daw ND, Dayan P. Choice values. Nat Neurosci. 2006;9:987–8.
31. Colas JT, Pauli WM, Larsen T, Tyszka JM, O'Doherty JP. Distinct prediction errors in mesostriatal circuits of the human brain mediate learning about the values of both states and actions: evidence from high-resolution fMRI. PLOS Computational Biol. 2017;13:e1005810.
32. Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, et al. Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. Arch Gen Psychiatry. 2012;69:129–38.
33. Hernaus D, Gold JM, Waltz JA, Frank MJ. Impaired expected value computations coupled with overreliance on stimulus-response learning in schizophrenia. Biol Psychiatry: Cogn Neurosci Neuroimaging. 2018;3:916–26.
34. Ghods-Sharifi S, Floresco SB. Differential effects on effort discounting induced by inactivations of the nucleus accumbens core or shell. Behav Neurosci. 2010;124:179–91.
35. Salamone JD, Correa M. The mysterious motivational functions of mesolimbic dopamine. Neuron.2012;76:470–85.
36. Hall J, Parkinson JA, Connor TM, Dickinson A, Everitt BJ. Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behaviour. Eur J Neurosci. 2001;13:1984–92.
37. Corbit LH, Balleine BW. The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. J Neurosci. 2011;31:11786–94.
38. Chib VS, De Martino B, Shimojo S, O'Doherty JP. Neural mechanisms underlying paradoxical performance for monetary incentives are driven by loss aversion. Neuron.2012;74:582–94.
39. Niv Y, Joel D, Dayan P. A normative perspective on motivation. Trends Cogn Sci. 2006;10:375–81.
40. Camille N, Tsuchida A, Fellows LK. Double dissociation of stimulus-value and action-value learning in humans with orbitofrontal or anterior cingulate cortex damage. J Neurosci. 2011;31:15048–52.
41. Ostlund SB, Balleine BW. Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. J Neurosci. 2007;27:4819–25.

42. Rudebeck PH, Behrens TE, Kennerley SW, Baxter MG, Buckley MJ, Walton ME, et al. Frontal cortex subregions play distinct roles in choices between actions and stimuli. J Neurosci. 2008;28:13775–85.
43. Rushworth MF, Behrens TEJ, Rudebeck PH, Walton ME. Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. Trends Cogn Sci. 2007;11:168–76.
44. Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE. Frontal cortex and reward-guided learning and decision-making. Neuron.2011;70:1054–69.
45. O'Doherty JP. Contributions of the ventromedial prefrontal cortex to goal-directed action selection. Ann N. Y Acad Sci. 2011;1239:118–29.
46. Padoa-Schioppa C, Assad JA. Neurons in the orbitofrontal cortex encode economic value. Nature. 2006;441:223–6.
47. Hyman JM, Whitman J, Emberly E, Woodward TS, Seamans JK. Action and outcome activity state patterns in the anterior cingulate cortex. Cereb Cortex. 2013;23:1257–68.
48. Platt ML, Glimcher PW. Neural correlates of decision variables in parietal cortex. Nature.1999;400:233–8.
49. Quilodran R, Rothe M, Procyk E. Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron.2008;57:314–25.
50. Sugrue LP, Corrado GS, Newsome WT. Matching behavior and the representation of value in the parietal cortex. Science.2004;304:1782–7.
51. Averbeck BB, Murray EA. Hypothalamic interactions with large-scale neural circuits underlying reinforcement learning and motivated behavior. Trends Neurosci. 2020;9:681–694.
52. Sternson SM. Hypothalamic survival circuits: blueprints for purposive behaviors. Neuron.2013;77:810–24.
53. Andersen RA, Snyder LH, Bradley DC, Xing J. Multimodal representation of space in the posterior parietal cortex and its use in planning movements. Annu Rev Neurosci. 1997;20:303–30.
54. Genovesio A, Wise SP, Passingham RE. Prefrontal–parietal function: from foraging to foresight. Trends Cogn Sci. 2014;18:72–81.
55. Lauwereyns J, Watanabe K, Coe B, Hikosaka O. A neural correlate of response bias in monkey caudate nucleus. Nature.2002;418:413–7.
56. Holroyd CB, Yeung N. An integrative theory of anterior cingulate cortex function: option selection in hierarchical reinforcement learning. In *Neural Basis of Motivational and Cognitive Control*, edited by R. B. Mars, J. Sallet, M. F. S. Rushworth, and N. Yeung. Cambridge, MA: MIT Press; 333–49.
57. Averbeck BB, Sohn J-W, Lee D. Activity in prefrontal cortex during dynamic selection of action sequences. Nat Neurosci. 2006;9:276–82.
58. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature.2006;441:876–9.
59. Schönberg T, Daw ND, Joel D, O'Doherty JP. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci. 2007;27:12860–7.
60. Wilson RC, Takahashi YK, Schoenbaum G, Niv Y. Orbitofrontal cortex as a cognitive map of task space. Neuron.2014;81:267–79.
61. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8:1704–11.
62. Dickinson A. Actions and habits: the development of behavioural autonomy. Philos Trans R Soc Lond B, Biol Sci. 1985;308:67–78.
63. Adams CD. Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q J Exp Psychol Sect B. 1982;34:77–98.
64. Balleine BW, Dickinson A. The role of incentive learning in instrumental outcome revaluation by sensory-specific satiety. Anim Learn Behav. 1998;26:46–59.
65. Yin HH, Knowlton BJ, Balleine BW. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci. 2004;19:181–9.
66. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci. 2005;22:513–23.
67. Rudebeck PH, Saunders RC, Prescott AT, Chau LS, Murray EA. Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat Neurosci. 2013;16:1140–5.
68. Reber J, Feinstein JS, O'Doherty JP, Liljeholm M, Adolphs R, Tranel D. Selective impairment of goal-directed decision-making following lesions to the human ventromedial prefrontal cortex. Brain.2017;140:1743–56.
69. Valentin VV, Dickinson A, O'Doherty JP. Determining the neural substrates of goal-directed learning in the human brain. J Neurosci. 2007;27:4019–26.
70. Balleine BW. The meaning of behavior: discriminating reflex and volition in the brain. Neuron.2019;104:47–62.
71. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron.2011;69:1204–15.
72. Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND. Model-based choices involve prospective neural activity. Nat Neurosci. 2015;18:767–72.
73. Huang Y, Yaple ZA, Yu R. Goal-oriented and habitual decisions: neural signatures of model-based and model-free learning. NeuroImage.2020;215:116834.
74. Kim D, Park GY, O'Doherty JP, Lee SW. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. Nat Commun. 2019;10:1–14.
75. Lee SW, Shimojo S, O'Doherty JP. Neural computations underlying arbitration between model-based and model-free learning. Neuron.2014;81:687–99.
76. Akam T, Costa R, Dayan P. Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. PLOS Computational Biol. 2015;11:e1004648.
77. Feher da Silva C, Hare TA. Humans primarily use model-based inference in the two-stage task. Nat Hum Behav. 2020; 4(10):1053–1066.
78. Collins AG, Cockburn J. Beyond dichotomies in reinforcement learning. Nat Rev Neurosci. 2020;21:576–86.
79. O'Doherty JP, Lee S, Tadayonnejad R, Cockburn J, Iigaya K, Charpentier CJ. Why and how the brain weights contributions from a mixture of experts. Neurosci Biobehav Rev. 2021;123:14–23.
80. Prevost C, McCabe JA, Jessup RK, Bossaerts P, O'Doherty JP. Differentiable contributions of human amygdalar subregions in the computations underlying reward and avoidance learning. Eur J Neurosci. 2011;34:134–45.
81. Pauli WM, Gentile G, Collette S, Tyszka JM, O'Doherty JP. Evidence for model-based encoding of Pavlovian contingencies in the human brain. Nat Commun. 2019;10:1099.
82. Pool ER, Pauli WM, Kress CS, O'Doherty JP. Behavioural evidence for parallel outcome-sensitive and outcome-insensitive Pavlovian learning systems in humans. Nat Hum Behav. 2019;3:284–96.
83. Averbeck BB. Theory of choice in bandit, information sampling and foraging tasks. PLOS Computational Biol. 2015;11:e1004164.
84. Botvinick MM, Niv Y, Barto AG. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition.2009;113:262–80.
85. Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, et al. A neural signature of hierarchical reinforcement learning. Neuron. 2011;71:370–9.
86. Badre D, D'Esposito M. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. J Cogn Neurosci. 2007;19:2082–99.
87. Badre D, Frank MJ. Mechanisms of hierarchical reinforcement learning in cortico–striatal circuits 2: evidence from fMRI. Cereb Cortex. 2012;22:527–36.
88. Koechlin E, Ody C, Kouneiher F. The architecture of cognitive control in the human prefrontal cortex. Science. 2003;302:1181–5.
89. Rhodes BJ, Bullock D, Verwey WB, Averbeck BB, Page MPA. Learning and production of movement sequences: behavioral, neurophysiological, and modeling perspectives. Hum Mov Sci. 2004;23:699–746.
90. Fujii N, Graybiel AM. Representation of action sequence boundaries by macaque prefrontal cortical neurons. Science. 2003;301:1246–9.
91. Martiros N, Burgess AA, Graybiel AM. Inversely active striatal projection neurons and interneurons selectively delimit useful behavioral sequences. Curr Biol. 2018;28:560–73.e5.
92. Averbeck BB, Lee D. Prefrontal neural correlates of memory for sequences. J Neurosci. 2007;27:2204–11.
93. Averbeck BB, Chafee MV, Crowe DA, Georgopoulos AP. Parallel processing of serial movements in prefrontal cortex. Proc Natl Acad Sci U.S.A. 2002;99:13172–7.
94. Tomov MS, Yagati S, Kumar A, Yang W, Gershman SJ. Discovery of hierarchical representations for efficient planning. PLOS Computational Biol. 2020;16: e1007594.
95. Schapiro AC, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM. Neural representations of events arise from temporal community structure. Nat Neurosci. 2013;16:486–92.
96. Dezfouli A, Balleine BW. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLOS Computational Biol. 2013;9:e1003364.
97. Shadlen MN, Newsome WT. Neural basis of a perceptual decision in the parietal cortex (Area LIP) of the Rhesus Monkey. J Neurophysiol. 2001;86:1916–36.
98. Hanks TD, Summerfield C. Perceptual decision making in rodents, monkeys, and humans. Neuron.2017;93:15–31.
99. Rangel A, Hare T. Neural computations associated with goal-directed choice. Curr Opin Neurobiol. 2010;20:262–70.
100. Fan Y, Gold JI, Ding L. Frontal eye field and caudate neurons make different contributions to reward-biased perceptual decisions. ELife. 2020;9:e60535.
101. Ratcliff R, McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural Comput. 2007;20:873–922.
102. Gold JI, Shadlen MN. Representation of a perceptual decision in developing oculomotor commands. Nature. 2000;404:390–4.
103. Hanks TD, Kopec CD, Brunton BW, Duan CA, Erlich JC, Brody CD. Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature.2015;520:220–3.

104. Krajbich I, Armel C, Rangel A. Visual fixations and the computation and comparison of value in simple choice. Nat Neurosci. 2010;13:1292–8.
105. Basten U, Biele G, Heekeren HR, Fiebach CJ. How the brain integrates costs and benefits during decision making. PNAS.2010;107:21767–72.
106. Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A. Transformation of stimulus value signals into motor commands during simple choice. PNAS.2011;108:18120–5.
107. Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG. A general mechanism for perceptual decision-making in the human brain. Nature.2004;431:859–62.
108. Polanía R, Krajbich I, Grueschow M, Ruff CC. Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. Neuron.2014;82:709–20.
109. Collins AG, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012;35:1024–35.
110. Beiser DG, Hua SE, Houk JC. Network models of the basal ganglia. Curr Opin Neurobiol. 1997;7:185–90.
111. Frank MJ, Claus ED. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychological Rev. 2006;113:300–26.
112. Alexander GE, Crutcher MD. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci. 1990;13:266–71.
113. DeLong MR. Primate models of movement disorders of basal ganglia origin. Trends Neurosci. 1990;13:281–5.
114. Cox J, Witten IB. Striatal circuits for reward learning and decision-making. Nat Rev Neurosci. 2019;20:482–94.
115. Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, et al. Concurrent activation of striatal direct and indirect pathways during action initiation. Nature.2013;494:238–42.
116. Klaus A, Martins GJ, Paixao VB, Zhou P, Paninski L, Costa RM. The spatiotemporal organization of the striatum encodes action space. Neuron.2017;95:1171–80.e7.
117. Donahue CH, Liu M, Kreitzer AC. Distinct value encoding in striatal direct and indirect pathways during adaptive learning. BioRxiv. 2018. https://doi.org/10.1101/277855.
118. Nonomura S, Nishizawa K, Sakai Y, Kawaguchi Y, Kato S, Uchigashima M, et al. Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. Neuron.2018;99:1302–14.e5.
119. Yttri EA, Dudman JT. Opponent and bidirectional control of movement velocity in the basal ganglia. Nature.2016;533:402–6.
120. Hikida T, Kimura K, Wada N, Funabiki K, Nakanishi S. Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. Neuron.2010;66:896–907.
121. Collins AG, Frank MJ. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychological Rev. 2014;121:337.
122. Yartsev MM, Hanks TD, Yoon AM, Brody CD. Causal contribution and dynamical encoding in the striatum during evidence accumulation. ELife.2018;7:e34929.
123. Jessup RK, O'Doherty JP. Human dorsal striatal activity during choice discriminates reinforcement learning behavior from the Gambler's Fallacy. J Neurosci. 2011;31:6296–304.
124. Houk JC, Davis JL, Beiser DG. Models of information processing in the Basal Ganglia. Cambridge, MA: MIT press; 1995,
125. Frank MJ. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci. 2005;17:51–72.
126. Thura D, Cisek P. The Basal Ganglia do not select reach targets but control the urgency of commitment. Neuron.2017;95:1160–70.e5.
127. Hikosaka O, Takikawa Y, Kawagoe R. Role of the basal ganglia in the control of purposive saccadic eye movements. Physiol Rev. 2000;80:953–78.
128. Roseberry TK, Lee AM, Lalive AL, Wilbrecht L, Bonci A, Kreitzer AC. Cell-type-specific control of brainstem locomotor circuits by Basal Ganglia. Cell.2016;164:526–37.
129. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. Humans use directed and random exploration to solve the explore–exploit dilemma. J Exp Psychol: Gen. 2014;143:2074–81.
130. Zajkowski WK, Kossut M, Wilson RC. A causal role for right frontopolar cortex in directed, but not random, exploration. ELife. 2017;6:e27430.
131. Ebitz RB, Albarran E, Moore T. Exploration disrupts choice-predictive signals and alters dynamics in prefrontal cortex. Neuron.2018;97:450–61.e9.
132. Wittmann BC, Daw ND, Seymour B, Dolan RJ. Striatal activity underlies novelty-based choice in humans. Neuron.2008;58:967–73.
133. Gustafson NJ, Daw ND. Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. PLoS Comput Biol. 2011;7:e1002235.
134. Redish AD. Vicarious trial and error. Nat Rev Neurosci. 2016;17:147.
135. Doll BB, Simon DA, Daw ND. The ubiquity of model-based reinforcement learning. Curr Opin Neurobiol. 2012;22:1075–81.
136. O'Keefe J The hippocampal cognitive map and navigational strategies. Brain and space, New York, NY, US: Oxford University Press; 1991. p. 273–95.
137. Miller KJ, Botvinick MM, Brody CD. Dorsal hippocampus contributes to model-based planning. Nat Neurosci. 2017;20:1269–76.
138. Vikbladh OM, Meager MR, King J, Blackmon K, Devinsky O, Shohamy D, et al. Hippocampal contributions to model-based planning and spatial memory. Neuron.2019;102:683–93.e4.
139. Stachenfeld KL, Botvinick MM, Gershman SJ. The hippocampus as a predictive map. Nat Neurosci. 2017;20:1643.
140. Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Computational Biol. 2017;13:e1005768.
141. Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ. The successor representation in human reinforcement learning. Nature Human. Behaviour.2017;1:680–92.
142. Constantinescu AO, O'Reilly JX, Behrens TEJ. Organizing conceptual knowledge in humans with a gridlike code. Science.2016;352:1464–8.
143. Tavares RM, Mendelsohn A, Grossman Y, Williams CH, Shapiro M, Trope Y, et al. A map for social navigation in the human brain. Neuron.2015;87:231–43.
144. Howard JD, Gottfried JA, Tobler PN, Kahnt T. Identity-specific coding of future rewards in the human orbitofrontal cortex. Proc Natl Acad Sci. 2015;112:5195–200.
145. Klein-Flügge MC, Barron HC, Brodersen KH, Dolan RJ, Behrens TEJ. Segregated encoding of reward–identity and stimulus–reward associations in human orbitofrontal cortex. J Neurosci. 2013;33:3202–11.
146. Kennerley SW, Behrens TE, Wallis JD. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat Neurosci. 2011;14:1581.
147. Wallis JD, Kennerley SW. Heterogeneous reward signals in prefrontal cortex. Curr Opin Neurobiol. 2010;20:191–8.
148. Suzuki S, Cross L, O'Doherty JP. Elucidating the underlying components of food valuation in the human orbitofrontal cortex. Nat Neurosci. 2017;20:1780–6.
149. O'Doherty JP, Rutishauser U, Iigaya K. The hierarchical construction of value. Curr Opin Behav Sci. 2021;41:71–7.
150. Butter CM. Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in Macaca mulatta. Physiol Behav. 1969;4:163–71.
151. Iversen SD, Mishkin M. Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. Exp Brain Res. 1970;11:376–86.
152. Dias R, Robbins TW, Roberts AC. Dissociation in prefrontal cortex of affective and attentional shifts. Nature.1996;380:69–72.
153. Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci. 2006;26:8360–7.
154. Jang AI, Costa VD, Rudebeck PH, Chudasama Y, Murray EA, Averbeck BB. The role of frontal cortical and medial-temporal lobe brain areas in learning a Bayesian prior belief on reversals. J Neurosci. 2015;35:11751–60.
155. Bartolo R, Averbeck BB. Prefrontal cortex predicts state switches during reversal learning. Neuron.2020;106:1044–.e4.
156. Costa VD, Tran VL, Turchi J, Averbeck BB. Reversal learning and dopamine: a bayesian perspective. J Neurosci. 2015;35:2407–16.
157. Starkweather CK, Gershman SJ, Uchida N. The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. Neuron.2018;98:616–29. e6.
158. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010;66:585–95.
159. Cross L, Cockburn J, Yue Y, O'Doherty JP. Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. Neuron. 2021;109(4), 724–738.
160. Dickinson A, Nicholas DJ, Adams CD. The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. Q J Exp Psychol Sect B. 1983;35:35–51.
161. Kool W, Gershman SJ, Cushman FA. Cost-benefit arbitration between multiple reinforcement-learning systems. Psychol Sci. 2017;28:1321–33.
162. Shenhav A, Botvinick MM, Cohen JD. The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron.2013;79:217–40.
163. Pezzulo G, Rigoli F, Chersi F. The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. Front Psychol. 2013;4:92. https://doi.org/10.3389/fpsyg.2013.00092.
164. Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol. 2011;7:e1002055.

165. Dromnelle R, Renaudo E, Pourcel G, Chatila R, Girard B, Khamassi M. How to reduce computation time while sparing performance during robot navigation? A neuro-inspired architecture for autonomous shifting between model-based and model-free learning. ArXiv:200414698 [Cs]. 2020.
166. Bogdanov M, Timmermann JE, Gläscher J, Hummel FC, Schwabe L. Causal role of the inferolateral prefrontal cortex in balancing goal-directed and habitual control of behavior. Sci Rep. 2018;8:9382.
167. Weissengruber S, Lee SW, O'Doherty JP, Ruff CC. Neurostimulation reveals context-dependent arbitration between model-based and model-free reinforcement learning. Cereb Cortex. 2019;29:4850–62.
168. Moran R, Keramati M, Dolan RJ. Model based planners reflect on their model-free propensities. PLOS Computational Biol. 2021;17:e1008552.
169. Frank MJ, Badre D. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. Cereb Cortex. 2012;22:509–26.
170. Doya K, Samejima K, Katagiri K, Kawato M. Multiple model-based reinforcement learning. Neural Comput. 2002;14:1347–69.
171. Dorfman HM, Gershman SJ. Controllability governs the balance between Pavlovian and instrumental action selection. Nat Commun. 2019;10:5826.
172. Charpentier CJ, Iigaya K, O'Doherty JP. A Neuro-computational account of Arbitration between choice imitation and goal emulation during human observational learning. Neuron. 2020;106(4):687–699.
173. Hamid AA, Frank MJ, Moore CI. Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. Cell. 2021;184(10):2733–2749.
174. O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. Ann N. Y Acad Sci. 2007;1104:35–53.
175. Marr D. Vision: a computational investigation into the human representation and processing of visual information. Cambridge,Mass:MIT press; 2010.
176. Márton CD, Schultz SR, Averbeck BB. Learning to select actions shapes recurrent dynamics in the corticostriatal system. Neural Netw. 2020;132:375–93.
177. Brown J, Bullock D, Grossberg S. How the Basal Ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. J Neurosci. 1999;19:10502–11.
178. McClelland JL, Rumelhart DE, University of California SD, PDP Research Group. Parallel distributed processing: explorations in the microstructure of cognition v. 2. Cambridge, Mass: MIT Press; 1986.
179. O'Reilly RC. Six principles for biologically based computational models of cortical cognition. Trends Cogn Sci. 1998;2:455–62.
180. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature.2015;521:436–44.
181. Tsuda B, Tye KM, Siegelmann HT, Sejnowski TJ. A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. PNAS. 2020;117:29872–82.
182. Yang GR, Joglekar MR, Song HF, Newsome WT, Wang X-J. Task representations in neural networks trained to perform many cognitive tasks. Nat Neurosci. 2019;22:297–306.
183. Gläscher JP, O'Doherty JP. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdiscip Rev: Cogn Sci. 2010;1:501–10.

AUTHOR CONTRIBUTIONS

BBA and JOD jointly planned the scope of the review, reviewed the literature, prepared figures, wrote the paper and revised the paper following reviewer comments.

FUNDING

JOD is supported by grants from the National Institutes of Mental Health (R01MH11425, R01MH121089, R21MH120805 and the NIMH Caltech Conte Center on the neurobiology of social decision-making, P50MH094258) and the National Institute on Drug Abuse (R01DA040011). This work was supported in part by the intramural research program of NIMH (BBA: ZIA MH002928).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to B.A. or J.P.O'D.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.