Actions, Policies, Values, and the Basal Ganglia

Nathaniel D. Daw, Yael Niv, and Peter Dayan

11th February 2005

Abstract

The basal ganglia are widely believed to be involved in the learned selection of actions. Building on this idea, reinforcement learning (RL) theories of optimal control have had some success in explaining the responses of their key dopaminergic afferents. While these *model-free* RL theories offer a compelling account of a range of neurophysiological and behavioural data, they offer only an incomplete picture of action control in the brain. Psychologists and behavioural neuroscientists have long appealed to the existence of at least two separate control systems underlying the learned control of behaviour. The *habit* system is closely identified with the basal ganglia, and we associate it with the model-free RL theories. The other system, more loosely localised in prefrontal regions and without such a detailed theoretical account, is associated with cognitively more sophisticated *goal-directed* actions. On the critical issue of which system determines the ultimate output when they disagree, there is a wide range of experimental results and sparse theoretical underpinning.

Here, we extend the RL account of neural action control by first interpreting goal-directed actions in terms of an alternative *model-based* strategy for RL. Then, by considering the relative *uncertainties* of model-free and model-based controllers, we offer a new and more comprehensive account of the confusing experimental results about how the systems trade off control. Our theory offers a more sharply delineated view of the contributions of the basal ganglia to learned behavioural control.

1 Introduction

The basal ganglia, and specifically the dorsal striatum, are widely believed to be involved in aspects of the learned selection of actions [70]. This belief has motivated, and in turn been reinforced by, a concerted effort to understand basal ganglia function in terms of theoretical ideas from adaptive optimal control, and particularly a branch of this called reinforcement learning (RL) [84]. Optimal control is particularly challenging, and therefore particularly interesting, in sequential tasks such as mazes, in which the rewarding or punishing consequences of a decision may take time to play out.

One method from RL, the temporal difference (TD) learning algorithm [83], in association with the so-called *actor-critic* model, has been specially important for neural modeling. This is because a key signal in

the algorithm, the TD prediction error, accurately captures many aspects of the phasic activity of midbrain dopamine neurons in animals performing appetitive tasks [66, 53, 79]. Amongst other places, these neurons deliver dopamine to the ventral and dorsal striatum and their afferents, where they likely control plasticity [89]. In this framework [66, 67], dopaminergically controlled plasticity in the ventral striatum and its afferents has been associated with the critic, which learns predictions of long term future reward. Dopaminergically controlled plasticity in the dorsal striatum has been associated with the actor, which specifies the subjects' action choices.

There are many algorithmic routes toward *optimal* behaviour. Insight into the actual structure of control is better provided by studying their characteristic forms of *sub*optimality. Indeed, a striking conjunction between theory and experiment arises when subjects are exposed to shifts in the circumstances or contingencies of a task, for instance, to motivational manipulations that devalue food reward by shifting the subject from hunger to satiation.[1] In this case, when the dorsal striatum appears to be in control of the choices of actions, the behaviour of the subjects is indeed suboptimal in that it fails to respect the change. Instead, subjects persist in performing actions that bring about rewards in which they are motivationally uninterested. A simple, so-called *model-free* form of the actor-critic has exactly the same problem, because it bases its decisions on *stored* predictions that require relearning for them to be revised in the light of the new values of the outcomes. This characteristic is not mere happenchance – rather it critically reflects the way the actor-critic uses such predictions to ameliorate the difficulty of choosing between actions whose consequences for reward or punishment might otherwise be obscure due to delay.

The results of numerous behavioural experiments (reviewed in [36, 37, 39]) employing such manipulations show that only some classes of behaviour are devaluation-insensitive in this manner, and thus that the actor-critic is at best an incomplete theory of action selection. Other behaviours react instantly to revaluation treatments, without the need for relearning. The two classes of behaviour are operationally defined as being devaluation-insensitive or *habitual* and devaluation-sensitive or *goal-directed* (since devaluations change a subject's goals). In some cases in which a sequence of multiple actions has to be performed to attain reward, experiments have even demonstrated behavioural inconsistencies, with subjects performing the first action in the chain, but then omitting the second action for a reward that is not desired.

The complete set of rather rococo behavioural distinctions has led psychologists to theories involving multiple parallel routes to action [36, 37, 39, 59], and has also inspired some modelling [30, 31]. However, there has not yet been a full computational investigation of the goal-directed route to action, or into the competition between these parallel routes.

In this chapter, we first consider goal-directed control in terms of *model-based* methods of RL. Rather than storing long-term values, these methods anticipate the long-term consequences of actions by searching in a *forward model* of the task contingencies. Along with other authors, we locate the major substrate of goal-directed control in the prefrontal cortex. Model-free and model-based controllers thus have distinct anatomical substrates and employ different strategies for choosing between actions whose consequences might take time to unfold. We then propose a theory of action choice that arbitrates between the two

[1] There are potentially important differences between devaluation achieved by such motivational shifts and devaluation achieved by, for instance, food aversion treatments. In this chapter, we suppress this distinction for simplicity.

controllers when their suggested actions disagree, proposing that competition for the control of behaviour is based on the relative reliability of the information upon which each system depends. This offers a new account for a body of seemingly puzzling behavioural and neuroscientific data, and a better delimited view of the contributions of the basal ganglia to learned behaviour.

2 Decision Strategies in Reinforcement Learning

In a typical RL setting, an agent is placed in an environment (such as a maze or an operant chamber), and allowed repeatedly to take actions (such as pressing a lever) and observe their consequences (such as gaining food pellets). The goal is defined in terms of learning to choose actions that optimise some long-term measure of future utility. As in a maze, where early decisions can have a huge impact on the speed with which the goal is ultimately attained, the chief difficulty is tracking the consequences of each action over an extended timescale.

The most straightforward strategy for optimal control is to build an internal *world model* detailing the immediate consequences of actions, and then to search this model recursively through many stages to anticipate and evaluate the long-term consequences of any particular choice. In psychological terms, such a model is a form of the cognitive map stressed by Tolman [85], and studied by many others. It consists of three components: transitions, outcomes, and utilities. The first defines how actions induce transitions between situations (called *states* of the world), *ie,* the probability of going from one state to another when a particular action is performed (going from room A to room B when turning right). The next is the mappings of states or actions to affectively important outcomes (such as the particular food available at some location in a maze). The last is the mapping from each outcome to its subjective affective utility, a scalar measure of desirability.[2] Crucially, this can change depending on the animal's state of deprivation (for instance, a food pellet will be more desirable when an animal is hungry than when it is full). A long-term measure of utility can simply be the sum, over a sequence of states, of the immediate utilities received in each.

Although building or learning a model is straightforward, using (*ie* searching) it to infer the best action is not. For instance, in a maze, in order to decide which action to take in one room, one has to use the world model to search through many different paths in an expanding tree of future possibilities, adding up expected utilities to determine the value expected for each path. Clearly, comparing the consequences of many long sequences of behaviour is practically impossible, and therefore an alternative strategy is essential. Starting as far back as the seminal contribution of Samuel [73], two main ideas have been suggested (distinct though often used together), namely *pruning* and *caching*. Pruning involves prioritising paths and exploring only a subset. Except in special cases, pruning introduces inaccuracy, since prizes or pitfalls may be left undiscovered.

The alternative to pruning is caching, *ie* storing anticipated results of the search. In the actor-critic, a temporal-difference algorithm [83] is used to learn cached *values* (long-term utilities) for each state. These

[2] RL treatments normally collapse the last two components, outcome identities and utilities, into a single function mapping states to *rewards*; we separate them to allow for a clearer treatment of motivational manipulation, and in the light of some lesion data [18].

can be used, in turn, to learn a cached *policy* indicating which action is best in each state [15, 82, 14]. [3] The policy itself can be represented in a number of ways; in this paper, we consider a version of the actor-critic called advantage learning [5], which specifies a policy by retrieving from storage exactly the quantities that tree search builds on the fly. Specifically, advantage learning caches quantities called "advantages" for each action in each state. The advantage of a particular action in a particular state represents the additional long-term value expected for taking that action over the baseline expectation reflected in the state's cached value; the sum of the state value and the advantage is the predicted long-term value for taking the action in the state.

Advantage learning, like other actor-critic methods, is *model-free* in that caching obviates the need for building and searching in a model. However, the key benefit from caching, that expansion of the search tree is unnecessary, is also its key limitation. If the task is changed, for instance by altering either the transitions (*eg*, blocking or opening routes in a maze, as studied by Tolman [85]) or the immediate utilities (*eg*, by shifting the motivational state of rats pressing a lever for food, from hungry to sated), then the cached values and policy will not change with it, unless there is an explicit opportunity for relearning of each relevant value (which typically requires multiple experiences of the entire trajectory from decision to outcome). By contrast, decisions derived from search in the full tree are immediately sensitive to changes in any of the transitions or utilities given only local experience with the changed aspects themselves, since the tree search will encounter and take into account the new contingencies.

Model-free and model-based controllers actually live at two ends of a spectrum. Elaborations of the sort explored in [30] embed some, but not all, features of a learned model into a model-free controller. It is also common to combine both caching and pruning, i.e. to substitute cached values for unexplored paths in partial tree search [73]. In RL it is conventional to explore such variants. However, we interpret the data on animal behavioural choice as suggesting that a different strategy is at work, involving the cooperation and competition of (at least) two separate and simultaneously active controllers, one model-free (and caching) and the other model-based (and perhaps using pruning). We discuss the substantial evidence implicating the basal ganglia in instantiating the model-free component, and the rather more flimsy evidence implicating the prefrontal cortex in the model-based component in section 4. RL theory does not offer extensive guidance as to how to combine multiple simultaneous controllers when they disagree. However, there is a body of work in areas such as multisensory integration arguing that combination should be based on relative accuracy or certainty [61, 93, 32, 44]. That is, circumstances that promote one system over the other should do so in virtue of their differential impact on the reliability of each system.

In the rest of this chapter, we first link particular forms of model-free (advantage learning) and model-based (tree-search) controllers to habitual and goal-directed action choice respectively. We then construct a theory of the uncertainties of the two sorts of controllers, and use it to account for, and re-interpret, animal behavioural data. Finally, we review the evidence about their neural realizations and the complexities of their interaction.

[3] Only the policy ultimately determines the decisions. However, in methods like actor-critic, it is essential to learn the values of states in order to learn the optimal policy.

3 Controller Competition

The difference between model-based and model-free controllers in the light of task changes mirrors venerable ideas in psychology about multiple routes to action (reviewed in [39]). As we mentioned, psychologists distinguish between habitual responses, characterised by insensitivity to manipulations of the outcome utility, and goal-directed actions, which are sensitive to such manipulations. That literature takes a somewhat subtly different view than we have described about the information relied on by each system. There, goal-directed actions are assumed to be controlled by associations between responses and their outcomes ('R-O' associations); habits are envisioned to be outcome-independent because they instead rely on associations between stimuli and responses ('S-R'). The RL quantities discussed above (policies, values, etc.) broadly parallel this associative learning terminology (with states in place of stimuli; actions for responses, etc.), but they are more general and precise. For instance, it is not clear how the concept of the R-O association applies when multiple actions and outcomes may occur sequentially or interleaved. This could be problematic, for instance, in extending to such situations a previous account of habit formation [36], which relies on the strength of R-O contingencies.

Both goal-directed and habitual action are forms of instrumental behaviour. We will also briefly discuss another class of behaviour, known as Pavlovian responses. Rather than being a chosen arbitrary action (such as a lever press), Pavlovian behaviours are thought to be emitted automatically when an animal anticipates reinforcement, as in the famous example of Pavlov's dog salivating in expectation of food. Pavlovian responses can compete with, and thereby complicate the analysis of, instrumental choices. As with instrumental choices, some Pavlovian responses seem to be sensitive to motivational manipulations, whilst others are insensitive (*eg* [50, 51]). Though we lack the space to develop the notion completely, we suggest that the same underlying computational framework applies to Pavlovian as well as to instrumental behaviour.

The difference between habitual and goal-directed controllers has led to a confusing wealth of experimental results on the way that actions are chosen when they disagree. In a typical such experiment (illustrated in Figure 1), an animal is placed in a deprived motivational state (*eg* hunger), and trained to perform an action such as pressing a lever, in order to obtain a motivationally relevant outcome such as food pellets. Normally, obtaining reward actually requires a sequence of actions: pressing the lever to release food, and pushing a flap on a food magazine to retrieve it. After training, the utility of the outcome is devalued. This can be accomplished, for instance, by feeding the animal until it is sated (a motivational shift), or by conditioning aversion to the food by pairing it with drug-induced illness. The animal is then tested to see whether it will still perform the actions associated with the devalued food. Importantly, this test is performed in extinction — without reward delivery — to ensure that any change in behaviour is attributable to existing knowledge about the outcome and its utility rather than to new experience with either.[4]

In some circumstances — especially for actions that have been moderately trained — devaluation indeed reduces performance of the action relative to non-devalued controls (Figure 1a,b) [2, 6]. Consistent with the folk-psychological maxim that repetition breeds habits, with increased training, goal-directed actions

[4]Of course, the reward omission itself causes a reduction in responding; to control for this, responding is compared to another group that is also tested in extinction but without having experienced devaluation of the outcome.

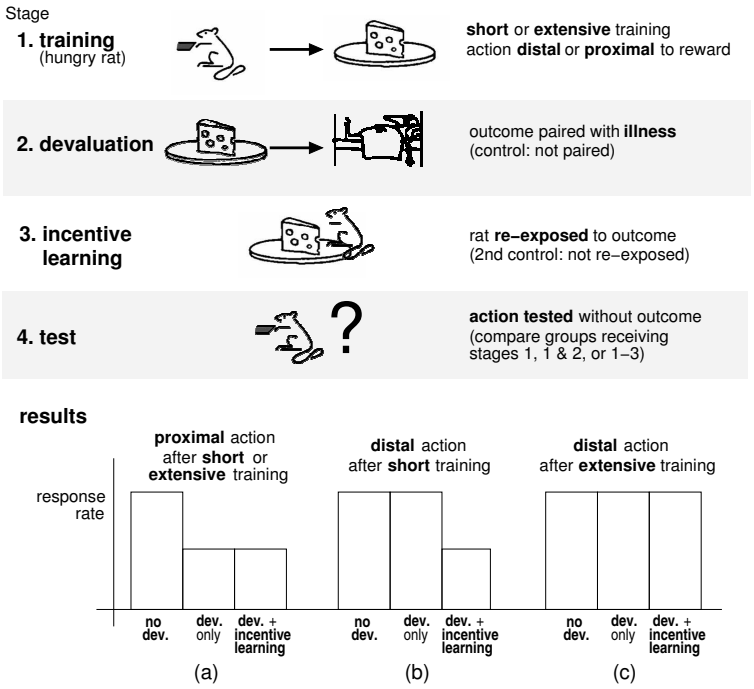

Figure 1: The stages in a typical devaluation experiment, together with a qualitative illustration of the results (based on data from, e.g., [12, 40]). Drawings from [38].

often become devaluation-insensitive, *ie*, habitual (Figure 1c) [1, 40]. However, there are a number of additional interacting factors, and hitherto no satisfactory explanation uniting them. For instance, magazine entries (the action more proximal to reward) can resist habitisation even while leverpresses become habitual [60] (Figure 1a). Devaluation sensitivity also persists despite overtraining for actions trained in more complex tasks involving multiple possible responses and outcomes [23, 49]. Table 1 summarises some of the important factors influencing devaluation sensitivity in experiments.

There is a further phenomenon that complicates interpreting these data with a two-controller model. If the reward is relatively novel (in particular, if it has never been experienced in the devalued state, *eg* while sated or after pairing with illness), then the outcome-sensitive actions subdivide further. Some actions are affected *immediately* by devaluation (Figure 1a), while other actions are affected by devaluation *only if* the animal is exposed to the outcome in the devalued state prior to the instrumental test [7, 6] (Figure 1b). This extra experience is called *incentive learning*, under the assumption (which we challenge below) that the re-exposure allows learning about the reduced utility ('incentive') of the outcome. Note that both of these categories differ from habits, which are usually devaluation-insensitive regardless of incentive learning [40] (Figure 1c). Remarkably, both profiles of devaluation sensitivity can be observed simultaneously in different behaviours occurring in sequence. For instance, in classic experiments by Balleine [6], magazine flap responses were immediately affected by devaluation, but contemporaneously recorded leverpresses were sensitive only after re-exposure. Similarly, in an experiment in which an animal had to carry out a sequence of two instrumental actions (a chain-pull and a leverpress) to obtain reinforcement, the proximal

devaluation sensitivity	devaluation insensitivity
moderate training	extensive training
multiple responses/outcomes	single response/outcome
incentive learning	no incentive learning
action proximal to reward	action distal from reward

Table 1: Some factors promoting devaluation sensitivity vs. insensitivity in behavioural experiments.

action was instantly devaluation-sensitive while the distal one required incentive learning [12].

We now offer an account of this apparent proliferation of behavioural categories, suggesting that they arise directly from the nature of the competition between the goal-directed and habitual controllers.

Uncertainty-based arbitration

As we have described it, decision-making within both the caching and tree-search systems depends on predicted action values. That is, the two systems employ *different* methods for estimating the *same* underlying quantities. The two estimation methods may therefore be more or less accurate under different circumstances. Here we propose that the experimental circumstances promoting devaluation sensitivity or insensitivity (Table 1) are those for which model search or caching, respectively, are relatively more accurate.

More concretely, we assume that arbitration between the two systems is based on the relative accuracy of their estimation methods. This raises the question of how these can be judged – it is known that such accuracies are extremely difficult to quantify [80, 57, 58]. One measure that arises naturally in many cases of competition in machine learning and neuroscience is *uncertainty* [61, 93, 32, 44]. Both systems' estimates of the action values have uncertainty (defined as the variances, *ie* expected squared errors, in the estimates) because the controllers have only limited experience on top of impoverished prior information.[5] We assume that the controllers themselves estimate the uncertainty in their action value predictions. This is computationally very difficult, but various approximations have been proposed [35, 34, 63]. The actor-critic can cache uncertainty information together with the cached values and advantages [35]. Instead of caching values, the model-based system stores an estimated world model, which is itself uncertain, and uses it to generate projected trajectories through the tree and accumulate action values over them. Uncertainty in the action values can be accumulated during this search based on the uncertainty in the estimates of the transitions, outcomes and utilities [34, 63]. This process is illustrated for a simple leverpressing task in Figure 2a, which shows how search proceeds forward through a sequence of anticipated states, and their values and uncertainties then propagate back up the tree.

[5]We distinguish *internal* from *external* uncertainty. External uncertainty results from actual stochasticity in the world (as when a lever-press pays off with 50% probability). Internal uncertainty is the Bayesian counterpart to the confidence interval in classical statistics, quantifying ignorance rather than stochasticity (as when an animal cannot reliably distinguish between payoff rates between 40% and 60% due to having only observed a few lever-presses). In differentially assessing the two controllers, it is internal uncertainty that is important.

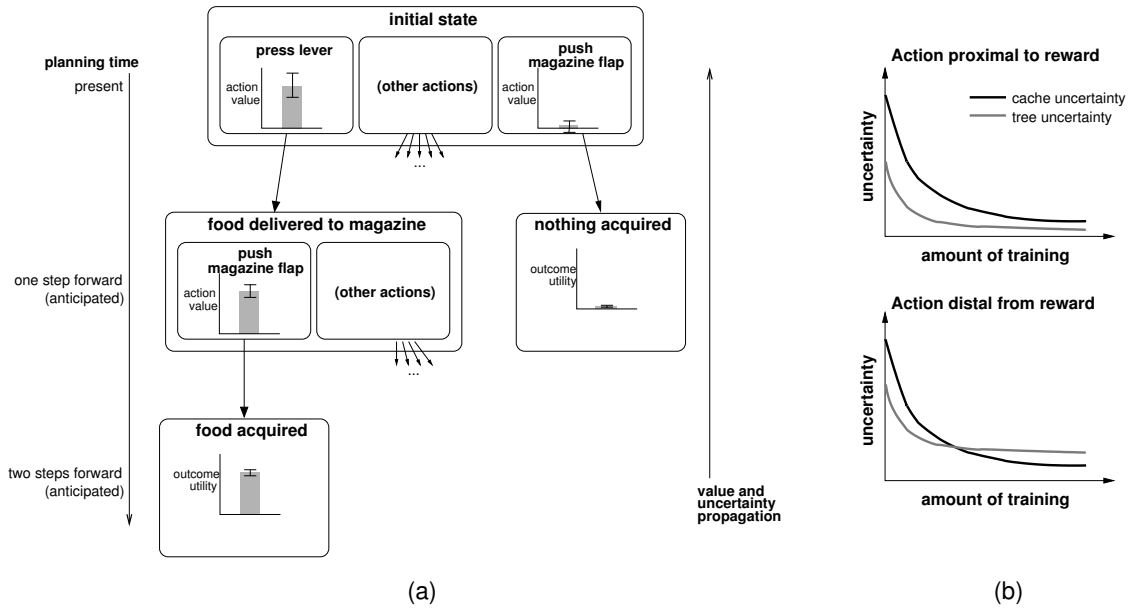

Figure 2: (a) Value and uncertainty propagation in tree search. Illustrated is a simple instrumental conditioning task in which a rat must press a lever and then push a flap, opening a food magazine to receive reward. Two paths through the tree (of many) are explored, and derived action values are shown with error bars representing uncertainty. Not directly depicted is uncertainty about the transitions (which state will result from an action), which is one reason that uncertainty about values grows moving up the tree, away from the rewards. (b) Uncertainty as a function of training for the cache and tree-search systems. Above, for actions proximal to reward, the tree is consistently more reliable than the cache, because it is more data-efficient, though the magnitude of this advantage declines with training. Below, for actions more distal from reward, the tree system is subject to additional uncertainty as a result of accumulated computational inaccuracy, allowing the cache system to surpass its reliability after some training.

The critical question for uncertainty-based competition is what makes one system more or less uncertain than the other. The tree based system stores all relevant information about past experience in its world model, and an exhaustive tree search can optimally incorporate all this information in constructing a value estimate. The caching system avoids searching in a world model, a shortcut that necessitates that new information be incorporated into its state estimates by a more haphazard and inefficient process of *bootstrapping*, or training each value estimate in terms of other estimates. Thus early in training (when data are scarce) cached values are only distantly related to their ultimate settings [80], and the more data-efficient tree system has an advantage in accuracy. This advantage is enhanced in more complex, data-intensive tasks, such as when there are multiple responses and outcomes, but wanes as experience accumulates (Figure 2b, top).

Why should the caching system ever dominate, if an *unlimited* tree-search system is strictly more accurate [57]? We assume that the tree search system has limitations of its own, for instance, that it is not capable of exhaustively searching all paths through the tree to derive accurate action values. In this case, it must rely on some approximation such as pruning (exploring only a subset of paths) or using a simplified approximation to the probability distribution over future states resulting from each contemplated action. This will introduce inaccuracy, which will compound over each step of search. Since the caching system requires no such approximation, its estimates would then be relatively advantaged for actions farther from the goal (*ie*, those that the goal-directed system must use deeper search to evaluate; Figure 2b, bottom). Which system is more reliable at evaluating any particular action then depends on the balance of the effects of inexperience and task complexity (favouring the tree system early in training and in complex tasks) against search depth (favouring the habit system for actions more distal from reward). Such a scheme unites and explains most of the experimental results concerning factors influencing outcome sensitivity, summarised in Table 1.

It remains for us to account for the effect of incentive learning. The key observation is that *uncertainty* in the habit system's value estimates is also cached, and thus (analogous to the value estimates themselves) insensitive to changes in confidence about the outcome utility. We assume that a devaluation without re-exposure immediately reduces the goal-directed system's utility estimate, and more importantly, for outcomes that have not been experienced in the new devalued state, that the change is also accompanied by increased uncertainty about the estimate of the new utility. Through the search process, the goal-directed system would immediately compute reduced values for actions leading to the devalued state, but also *increased uncertainties* about those values. This introduces an additional factor favouring the habit system, since because of caching, its uncertainties are (incorrectly) unaffected by devaluation. For actions that are only weakly under goal directed control (*eg*, more distal ones), devaluation without re-exposure can thus tip the balance in favour of the habit system. Performance of these actions would then be insensitive to devaluation, until a re-exposure treatment restored certainty in the goal-directed system's utility estimate, allowing it to reassert control. In contrast, the goal-directed system would retain control of those actions it most staunchly dominates (such as those more proximal to reward, or in tasks with multiple actions and outcomes) allowing them to adjust to devaluation without need for re-exposure. This pattern is consistent with the experimental data, explaining the difference between distal and proximal responsivity patterns depicted in Figure 1a,b [12] and similar differences in the effect of incentive learning when there are multiple versus single actions and outcomes [8, 72].

4 The Neural Substrate

We have so far concentrated mainly on a psychological picture of the controllers and their interaction. We now turn to the neural substrate, first of the two separate systems, and then their arbitration.

The core data anchoring the actor-critic model in the basal ganglia are recordings of midbrain dopamine neurons in primates engaged in appetitive learning tasks [76, 77]. The phasic responses of these neurons — to unpredicted but not predicted primary rewards, for example, and to cues signaling reward — closely resemble the TD error signal for reward prediction, which is used in the actor-critic to learn both values and policies. The correspondence between the TD signal and the neuronal responses has been worked out in detail [53, 66, 79, 56, 27]. Expanding from this identification of dopamine as a teaching signal, the actor-critic model proposes that dopamine targets — particularly in the striatum — are the sites of value and policy learning [53, 66, 81, 30]. Although much debated [52, 87], there is evidence suggesting that phasic dopamine is involved in appetitive but not aversive learning [65]. It has been, however, suggested that dorsal raphe serotonin complements this function in aversive learning, playing a similar role for predictions of punishment [28].

The separation between the two learning functions in the actor-critic parallels a fundamental division in the functional anatomy of the basal ganglia. Although there are many finer distinctions, a division of the striatum into dorsal and ventral subregions is well established anatomically, pharmacologically, physiologically, and even in functional neuroimaging. (For extensive references, see [88], who propose that the distinction is more accurately described as dorsolateral versus ventromedial.) The ventral striatum receives (and indirectly reciprocates) projections from limbic structures such as orbitofrontal cortex, and is implicated in reward and motivation [19]. Much of the dorsal striatum, by contrast, is connected with motor cortices and is itself implicated in learned motor control [3, 70]. Dopamine neurons are analogously grouped, into the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) [55].

The actor-critic model thus suggests that the dopamine projection from VTA, which targets ventral striatum and other limbic areas, supports the learning of state values, while that from SNc to dorsal striatum supports policy learning (Figure 3). An early observation influential in this identification is the little apparent difference between the activities of VTA and SNc dopamine neurons [78], despite their association with functionally quite disparate striatal territories. This negative finding is expected under the actor-critic, since values and policies share a common TD error signal. This overall functional mapping recently received more direct support in rodent neurophysiological [26] and human fMRI experiments [67] explicitly designed to exercise the two learning functions differentially. This proposal is also more consistent with anatomical data reviewed recently by Joel et al. [54] than is an alternative proposed mapping of actor and critic functions to different striatal subterritories [53].

Lesions of the dorsolateral striatum seem to abolish habitual control [92], as would be expected under the model. That is, even extensively trained actions show devaluation sensitivity in animals with such lesions. According to the model, the values cached in the *ventral* striatum are associated with states (and not actions), and so they would also be appropriate to drive Pavlovian responses (which typically anticipate

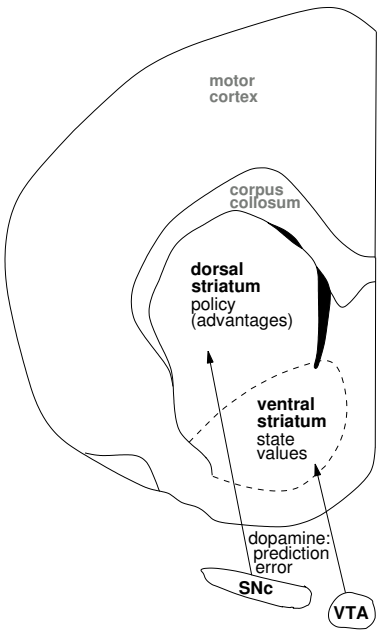

Figure 3: Suggested anatomical substrates of actor/critic model in midbrain and striatum.[6]

outcomes whose delivery is dependent on the state but not the animal's actions) as well as to support instrumental learning more dorsally. Indeed, lesions implicate the ventral striatum in some Pavlovian responses and in Pavlovian-instrumental interactions [47, 20]. Though this has been spottily tested, under our model, these are expected to be the devaluation-*insensitive* aspects of Pavlovian conditioning, since they originate from cached values. (As with instrumental choices, we expect devaluation-*sensitive* Pavlovian responses to originate in a model-based system.)

In contrast to the rather well specified habitual controller, there is much less data on the nature and substrates of outcome-sensitive, goal-directed actions [36, 37]. Broadly, one would expect this sort of planning function to be associated with prefrontal cortex, for a host of neuropsychological and neurophysiological reasons [69, 64, 21]. This general impression is supported by more specific studies of instrumental behaviour: Lesions of rat prelimbic cortex (a subarea of medial prefrontal cortex) abolish devaluation sensitivity for undertrained behaviours that would normally be goal-directed [10, 24, 60]. The full pattern of results suggests involvement of this area in representation or evaluation of the world model.

As described in Section 2, model-based planning also relies directly on information about outcomes and their utilities. Electrophysiological, lesion and imaging data implicate areas such as the orbitofrontal region of prefrontal cortex [75, 86, 68], gustatory insular cortex [10, 11], and basolateral amygdala [75, 17] in learning and representing these. Notably, lesions in all of these areas also abolish devaluation sensitivity [10, 45, 11, 13, 18], and further behavioural assays (*eg*, [17, 18]) support the interpretation that, in terms of our theory, these areas are involved in the mapping from states to outcomes, or outcomes to utilities, rather

[6]Figure adapted from *The Rat Brain in Stereotaxic Coordinates, Compact Third Editon*, Paxinos and Watson, copyright 1996, reprinted with permission from Elsevier.

than in the tree itself.

The lesion results we have discussed have broader implications for the organisation of control, in that they demonstrate the apparent possibility of *separately* disabling either the goal-directed or habitual controller, leaving the other intact. The results suggest a measure of independence between the systems, and indicate that both systems are capable of specifying actions even when they would not normally be in control (for instance that the habit system can control even moderately trained behaviour if the goal-directed system is disabled).

There is no direct evidence as to the neural basis of arbitration between the systems, though there have been suggestions that this function may be integrated into either system rather than being anatomically separate. Thus Redgrave and colleagues [71] propose that the basal ganglia may select between control systems (as well as between habitual actions). The converse view is that the prefrontal planning system performs the arbitration, by overriding or gating habitual control. Killcross and Coutereau [60] deploy this suggestion to explain their finding (paradoxical under the conventional anatomical mapping that we have described) that lesions of rat infralimbic cortex (another subarea of medial prefrontal cortex) disrupt control by the habit system, rather than by the goal-directed system [60, 25].

5 Discussion

We have offered a new framework for interpreting behavioural and physiological evidence regarding multiple routes to action, which contacts and enriches theories of RL and its neural substrates. We presented a bipartite model of action control, in which the predominant actor-critic model of dopaminergic and basal ganglia function (which is seen as the model of habitual control) is accompanied by a second, cortical system, which is capable of learning a model of the world and searching in it to plan actions. When the controllers differ in the actions they prefer, we have suggested that their relative certainties control which action gets executed, and have used this to explain the wealth of data as to circumstances promoting or inhibiting devaluation sensitivity, including the puzzling findings of incentive learning. Also, unlike some previous models in which goal-directed and habitual control are rather entangled [37, 30], the new model is consonant with lesion studies that suggest that the two behavioural systems are relatively dissociable and independent [60, 25, 92].

The standard account of incentive learning is different from the one we have presented. We view the goal-directed system as always being capable of immediate devaluation. The need for further learning arises because re-exposure restores control to this system over the habitual system. By contrast, the standard view [6, 9, 39] holds that the relevant actions are under the control of the goal-directed system throughout the incentive learning treatment, with re-exposure changing the value ascribed within this system. Immediate devaluation sensitivity for magazine entries is then explained by assuming that these are actually not controlled by the goal-directed or habitual systems, but are instead Pavlovian responses, controlled by a *separate,* Pavlovian evaluator whose utilities are immediately devaluable [41]. A main feature of our theory is the elimination of this appeal to a third system. Our account of these data is thus more parsimonious,

and also more directly applicable to Balleine et al.'s [12] demonstration of immediate devaluation even for unambiguously instrumental responses.

Various experimental studies could arbitrate between the two views of these issues. Notably, because we view non-devaluability prior to incentive learning as resulting from habitual rather than goal-directed control, we predict that lesions disabling the habit system [60] should render instrumental actions immediately susceptible to devaluation, eliminating the need for re-exposure. (The standard view predicts no such effect.) Further, in our view, the crucial aspect of re-exposure in an incentive learning experiment is its effect on *uncertainty* about the utility of an outcome rather than, as normally assumed, on the estimate of the utility itself. This may explain why incentive learning is dopamine-independent [43] — other, Pavlovian experiments testing manipulations of uncertainty instead implicate another neuromodulator, acetylcholine [48, 33]. We suggest that some of the same brain areas implicated in those studies may also play a role in incentive learning.

The model also has testable ramifications for the effects of devaluation on the responses of dopamine neurons. Dopamine responses to cues originate in the cached values, and should thus be insensitive to post-training changes in outcome value (*eg*, due to shifts in deprivation state). In contrast, dopamine responses to primary reward may be directly sensitive to devaluation, since they are based on the reward's observed utility. Neither of these predictions has been directly tested. However, the magnitude of the dopamine response to a cue has been shown to be correlated with the latency of a behavioural response triggered by that cue [74], and this has been taken as evidence that the dopamine responses carry motivational information. Our theory would indeed expect correlations between trial-to-trial fluctuations in dopamine responding and behavioural vigour (for habitual actions), since in advantage learning, both are controlled by the cached values. What has not been established is what systematic factors (if any) drove the fluctuations measured in the experiment [74]; one possibility is that the dopamine response is modulated by the animal's overall *drive*, which is a more generalised form of motivation that is *independent* of outcome expectancy and thus could be reported even by a system ignorant of the identity or utility of the outcome.

In common with other theories, our account of the goal-directed system is sketchy. The existing data are radically insufficient to verify whether devaluation-sensitive behaviours are actually products of a full, model-based search system, due to the extreme simplicity of the tasks used. There are many different algorithms for model learning and planning, and many shortcuts or hybrids that have advantages over either the actor/critic or full model search. Some of the earliest work on RL suggested that pruning and caching be combined, with cached values standing in for unexplored subtrees [73]; more recently, uncertainty has been used to decide where to make this substitution [16]. Applied to the model, these ideas would suggest a rich interaction between the goal-directed and habitual controllers — perhaps richer than the relative independence suggested by the lesion studies. Another hybrid is the successor representation [29], which was used in a previous RL model of goal-directed action [30]. This algorithm uses caching, but of expected *outcome identities* rather than policies or values. Because of this, it can adapt immediately to outcome devaluations such as those in the experiments discussed here, but not to manipulations changing the transition structure of the task (as by rearranging walls in a maze) that would invalidate the cached outcomes. The latter sort of manipulation was pioneered by Tolman [85], but has seen little testing in modern times. Understanding the

capabilities and limitations of the goal-directed system will require much more experimentation studying how organisms can react to different sorts of contingency changes in more complicated sequential decision tasks.

Also, we are at the earliest stages of understanding the factors governing uncertainty in the two systems and the way this influences arbitration. For instance, an alternative to our suggestion that inaccuracy accumulates due to approximate tree search is that whether exact or approximate, tree search incurs costs of energetic or computational resources (again accumulating over each step). This could, for deeper searches, outweigh the cost of potentially reduced utility that would result from abandoning the search altogether and falling back on the less accurate caching system. A distinct possibility for differential uncertainty that would favour the cache is that the two systems might be tuned to differing expectations about how quickly the estimated aspects of the world (transitions, utilities) are *changing*. The possibility of change is an ongoing source of uncertainty, and a system that assumed faster change would be more uncertain asymptotically but quicker to adapt in situations when the world actually changed (as for the Kalman filter, [4]). The goal-directed system, whose comparative advantage is most apparent when there is only little data, could have a built-in bias for transience, and therefore high asymptotic uncertainties, whereas the habit system, which anyway requires more substantial training, might have a built-in bias for stability and correspondingly lower asymptotic uncertainties. Finally, although we have spoken as though arbitration is all-or-nothing, it is a straightforward and reasonable extension to assume behaviour is based on both controllers' predicted values, with contributions weighted by their uncertainty.

Another crucial issue for future work is connecting the well-understood theory of discrete, sequential decisions (which we have drawn on here as an idealisation) with the realistic experimental situation, in which animals are free to emit responses in continuous time and analysis typically focuses on the *rates* with which they do so. A more careful treatment of these issues will be required before we can address a further factor known to influence devaluation, which has to do with how payoffs are scheduled as a function of responding (interval vs. ratio schedules, [42]). This issue is, at present, better explained under a previous account of habit formation [36].

Though we have not specifically discussed the circumstances under which Pavlovian responses should be sensitive to devaluation [50, 51], we expect the same considerations to apply as for instrumental action. The predictions triggering Pavlovian responses, like those driving choices, might originate either from model search or from cached values, with corresponding sensitivity or insensitivity to outcome devaluation. By eliminating the appeal to a separate and qualitatively different Pavlovian motivational system, our model opens the possibility of a unified account of Pavlovian and instrumental behaviours, in which both share a common forward model and common cached values. A major task for the future is to appraise this model in the light of the (equally complex) data on devaluation sensitivity in Pavlovian conditioning. Do the same factors impact Pavlovian as instrumental devaluation? Are the systems really shared? And if not, how do they interact? Of special relevance will be two paradigms that couple Pavlovian and instrumental conditioning (and also involve dopamine and ventral striatum), Pavlovian-instrumental transfer [62] and conditioned reinforcement [90].

Finally, what of the basal ganglia? We suggest that the new theory delimits more precisely the role of the

dorsal striatum in habitual behaviour, and also the roles of dopamine and of the ventral striatum's value predictions in learning those habits. We have not discussed dopamine's additional on-line, energising effect on behaviour (as in Pavlovian-instrumental transfer [43]). This likely also involves values learned in the ventral striatum, if only through their influence over the dopamine projection to the dorsal habit areas [46]. Rounding out the cornucopia of open concerns is the observation that the evaluation of search trees in the goal-directed system may tax working memory [24]. This suggests the unexplored possibility that even the goal-directed system may be affected by dopamine (and by striatal influences on it), which is thought to have a specific role in prefrontal mnemonic function [91, 22].

Acknowledgements

We are grateful to Tony Dickinson, Peter Holland, Daphna Joel, and Maneesh Sahani for helpful discussions, and to Rudolf Cardinal, Aaron Courville, Daphna Joel, and Genela Morris for thoughtful comments on an earlier draft. The authors are supported by the Gatsby Foundation, the EU BIBA project (PD), a Royal Society USA Research Fellowship (ND), and the Interdisciplinary Center for Neural Computation (YN).

References

[1] C. D. Adams. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 34B:77–98, 1982.

[2] C. D. Adams and A. Dickinson. Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 33B:109–112, 1981.

[3] G. E. Alexander, M. R. Delong, and P. L. Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9:357–381, 1986.

[4] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.

[5] L. C. Baird. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of the International Conference on Neural Networks, Orlando, FL*, 1994.

[6] B. Balleine. Instrumental performance following a shift in primary motivation depends upon incentive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 18:236–260, 1992.

[7] B. Balleine and A. Dickinson. Instrumental performance following reinforcer devaluation depends upon incentive learning. *Quarterly Journal of Experimental Psychology*, 43B:279–296, 1991.

[8] B. Balleine and A. Dickinson. Signalling and incentive processes in instrumental reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 45B:285–301, 1992.

[9] B. W. Balleine. Incentive processes in instrumental conditioning. In R. R. Mowrer and S. B. Klein, editors, *Handbook of Contemporary Learning Theories*, pages 307–366. Lawrence Erlbaum, Mahwah, NJ, 2000.

[10] B. W. Balleine and A. Dickinson. Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37:407–419, 1998.

[11] B. W. Balleine and A. Dickinson. The effect of lesions of the insular cortex on instrumental conditioning: Evidence for a role in incentive memory. *Journal of Neuroscience*, 20:8954–8964, 2000.

[12] B. W. Balleine, C. Garner, F. Gonzalez, and A. Dickinson. Motivational control of heterogeneous instrumental chains. *Journal of Experimental Psychology: Animal Behavior Processes*, 21:203–217, 1995.

[13] B. W. Balleine, A. S. Killcross, and A. Dickinson. The effect of lesions of the basolateral amygdala on instrumental conditioning. *Journal of Neuroscience*, 23:666–675, 2003.

[14] A. G. Barto. Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 215–232. MIT Press, Cambridge, MA, 1995.

[15] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuron-like adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13:834–46, 1983.

[16] E. B. Baum and W. D. Smith. A Bayesian approach to relevance in game playing. *Artificial Intelligence*, 97:195–242, 1997.

[17] P. Blundell, G. Hall, and S. Killcross. Lesions of the basolateral amygdala disrupt selective aspects of reinforcer representation in rats. *Journal of Neuroscience*, 21:9018–9026, 2001.

[18] P. Blundell, G. Hall, and S. Killcross. Preserved sensitivity to outcome value after lesions of the basolateral amygdala. *Journal of Neuroscience*, 23:7702–7709, 2003.

[19] R. N. Cardinal, J. A. Parkinson, J. Hall, and B. J. Everitt. Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews*, 26:321–352, 2002.

[20] R. N. Cardinal, J. A. Parkinson, G. Lachenal, K. M. Halkerston, N. Rudarakanchana, J. Hall, C. H. Morrison, S. R. Howes, T. W. Robbins, and B. J. Everitt. Effects of selective excitotoxic lesions of the nucleus accumbens core, anterior cingulate cortex, and central nucleus of the amygdala on autoshaping performance in rats. *Behavioral Neuroscience*, 116:553–567, 2002.

[21] L. Clark, R. Cools, and T. W. Robbins. The neuropsychology of ventral prefrontal cortex: Decision-making and reversal learning. *Brain and Cognition*, 55:41–53, 2004.

[22] J. D. Cohen, T. S. Braver, and J. W. Brown. Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology*, 12:223–229, 2002.

[23] R. M. Colwill and R. A. Rescorla. Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *Journal of Experimental Psychology: Animal Behavior Processes*, 11:520–526, 1985.

[24] L. H. Corbit and B. W. Balleine. The role of prelimbic cortex in instrumental conditioning. *Behavioral Brain Research*, 146:145–157, 2003.

[25] E. Coutureau and S. Killcross. Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioral Brain Research*, 146:167–174, 2003.

[26] N. D. Daw. *Reinforcement Learning Models of the Dopamine System and Their Behavioral Implications*. PhD thesis, School of Computer Science, Carnegie Mellon University, 2003.

[27] N. D. Daw, A. C. Courville, and D. S. Touretzky. Timing and partial observability in the dopamine system. In *Advances in Neural Information Processing Systems 15*, pages 99–106, Cambridge, MA, 2003. MIT Press.

[28] N. D. Daw, S. Kakade, and P. Dayan. Opponent interactions between serotonin and dopamine. *Neural Networks*, 15:603–616, 2002.

[29] P. Dayan. Improving generalisation for temporal difference learning: The successor representation. *Neural Computation*, 5:613–624, 1993.

[30] P. Dayan. Motivated reinforcement learning. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 11–18, Cambridge, MA, 2002. MIT Press.

[31] P. Dayan and B. W. Balleine. Reward, motivation and reinforcement learning. *Neuron*, 36:285–298, 2002.

[32] P. Dayan, S. Kakade, and P. R. Montague. Learning and selective attention. *Nature Neuroscience*, 3:1218–1223, 2000.

[33] P. Dayan and A. Yu. ACh, uncertainty, and cortical inference. In *Advances in Neural Information Processing Systems 14*, pages 189–196, 2001.

[34] R. Dearden, N. Friedman, and D. Andre. Model based Bayesian exploration. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 150–159, 1999.

[35] R. Dearden, N. Friedman, and S. J. Russell. Bayesian Q-learning. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI)*, pages 761–768, 1998.

[36] A. Dickinson. Actions and habits — the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B — Biological Sciences*, 308:67–78, 1985.

[37] A. Dickinson. Instrumental conditioning. In N. J. Mackintosh, editor, *Animal Learning and Cognition*, pages 45–79. Academic Press, San Diego, 1994.

[38] A. Dickinson and B. Balleine. Motivational control of goal-directed action. *Animal Learning and Behavior*, 22:1–18, 1994.

[39] A. Dickinson and B. Balleine. The role of learning in motivation. In C. R. Gallistel, editor, *Stevens' Handbook of Experimental Psychology (3rd ed.) Vol. 3: Learning, Motivation and Emotion*. Wiley, New York, 2002.

[40] A. Dickinson, B. Balleine, A. Watt, F. Gonzalez, and R. A. Boakes. Motivational control after extended instrumental training. *Animal Learning and Behavior*, 23:197–206, 1995.

[41] A. Dickinson and G. R. Dawson. Pavlovian processes in the motivational control of instrumental performance. *Quarterly Journal of Experimental Psychology*, 39B:201–213, 1987.

[42] A. Dickinson, D. J. Nicholson, and C. D. Adams. The effect of the instrumental contingency on susceptibility to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 35B:249–263, 1983.

[43] A. Dickinson, J. Smith, and J. Mirenowicz. Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behavioral Neuroscience*, 114:468–483, 2000.

[44] M.O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415:429–433, 2002.

[45] M. Gallagher, R. W. McMahan, and G. Schoenbaum. Orbitofrontal cortex and representation of incentive value in associative learning. *Journal of Neuroscience*, 19:6610–6614, 1999.

[46] S. N. Haber, J. L. Fudge, and N. R. McFarland. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, 20:2369–2382, 2000.

[47] J. Hall, J. A. Parkinson, T. M. Connor, A. Dickinson, and B. J. Everitt. Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behaviour. *European Journal of Neuroscience*, 13:1984–1992, 2001.

[48] P. C. Holland. Brain mechanisms for changes in processing of conditioned stimuli in Pavlovian conditioning: Implications for behavior theory. *Animal Learning and Behavior*, 25:373–399, 1997.

[49] P. C. Holland. Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, 30:104–117, 2004.

[50] P. C. Holland and R. A. Rescorla. The effect of two ways of devaluing the unconditioned stimulus after first- and second-order appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, 1:355–363, 1975.

[51] P. C. Holland and J. J. Straub. Differential effect of two ways of devaluing the unconditioned stimulus after Pavlovian appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, 5:65–78, 1979.

[52] J. C. Horvitz. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96:651–656, 2000.

[53] J. C. Houk, J. L. Adams, and A. G. Barto. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 249–270. MIT Press, Cambridge, MA, 1995.

[54] D. Joel, Y. Niv, and E. Ruppin. Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, 15:535–547, 2002.

[55] D. Joel and I. Weiner. The connections of the dopaminergic system with the striatum in rats and primates: An analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, 96:451–474, 2000.

[56] S. Kakade and P. Dayan. Dopamine: Generalization and bonuses. *Neural Networks*, 15:549–559, 2002.

[57] M. Kearns and S. Singh. Finite-sample rates of convergence for Q-learning and indirect methods. In *Advances in Neural Information Processing Systems 11*, pages 996–1002, 1999.

[58] M. Kearns and S. Singh. Bias-variance error bounds for temporal difference updates. In *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pages 142–147, 2000.

[59] S. Kilcross and P. Blundell. Associative representations of emotionally significant outcomes. In S. Moore and M. Oaksford, editors, *Emotional Cognition: From Brain to Behaviour*, pages 35–73. John Benjamins, Amsterdam, 2002.

[60] S. Killcross and E. Coutureau. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, 13:400–408, 2003.

[61] D. V. Lindley. Reconciliation of discrete probability distributions. In J. M. Bernado, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 375–390. Elsevier Science Publishers B.V., Amsterdam, 1985.

[62] P. F. Lovibond. Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, 9:225–247, 1983.

[63] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 2004. (in press).

[64] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001.

[65] J. Mirenowicz and W. Schultz. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, 379:449–451, 1996.

[66] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16:1936–1947, 1996.

[67] J. O'Doherty, P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R. J. Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304:452–454, 2004.

[68] J. O'Doherty, E. T. Rolls, S. Francis, R. Bowtell, F. McGlone, G. Kobal, B. Renner, and G. Ahne. Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex. *Neuroreport*, 11:893–897, 2000.

[69] A. M. Owen. Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Progress in Neurobiology*, 53:431–450, 1997.

[70] M. G. Packard and B. J. Knowlton. Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, 25:563–593, 2002.

[71] P. Redgrave, T. J. Prescott, and K. Gurney. The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023, 1999.

[72] R. A. Rescorla. A note on depression of instrumental responding after one trial of outcome devaluation. *Quarterly Journal of Experimental Psychology*, 47B:27–37, 1994.

[73] A. L. Samuels. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3:210–229, 1959.

[74] T. Satoh, S. Nakai, T. Sato, and Minoru Kimura. Correlated coding of motivation and outcome of decision by dopamine neurons. *Journal of Neuroscience*, 23:9913–9923, 2003.

[75] G. Schoenbaum, A. A. Chiba, and M. Gallagher. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience*, 1:155–159, 1998.

[76] W. Schultz. Activity of dopamine neurons in the behaving primate. *Seminars in the Neurosciences*, 4:129–138, 1992.

[77] W. Schultz. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27, 1998.

[78] W. Schultz, P. Apicella, and T. Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, 13:900–913, 1993.

[79] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997.

[80] S. P. Singh and P. Dayan. Analytical mean squared error curves in Temporal Difference learning. *Machine Learning*, 32:5–40, 1998.

[81] R. E. Suri and W. Schultz. Temporal difference model reproduces predictive neural activity. *Neural Computation*, 13:841–862, 2001.

[82] R. S. Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts, Amherst, MA, 1984.

[83] R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44, 1988.

[84] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[85] E. C. Tolman. *Purposive Behavior in Animals and Men*. Appleton-Century-Crofts, New York, 1932.

[86] L. Tremblay and W. Schultz. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398:704–708, 1999.

[87] M. A. Ungless, P. J. Magill, and J. P. Bolam. Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, 303:2040–2042, 2004.

[88] P. Voorn, L. J. Vanderschuren, H. J. Groenewegen, T. W. Robbins, and C. M. Pennartz. Putting a spin on the dorsal-ventral divide of the striatum. *Trends in Neuroscience*, 27:468–474, 2004.

[89] J. R. Wickens, A. J. Begg, and G. W. Arbuthnott. Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Neuroscience*, 70:1–5, 1996.

[90] B. A. Williams and R. Dunn. Conditioned reinforcement: Neglected or outmoded explanatory construct? *Psychonomic Bulletin and Review*, 1:457–475, 1994.

[91] G. V. Williams and P. S. Goldman-Rakic. Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature*, 376:572–575, 1995.

[92] H. H. Yin, B. J. Knowlton, and B. W. Balleine. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19:181–189, 2004.

[93] A. L. Yuille and H. H. Bültoff. Bayesian decision theory and psychophysics. In D. C. Knill and W. Richards, editors, *Perception as Bayesian Inference*, pages 123–161. Cambridge University Press, New York, 1996.