Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Reinforcement learning: The Good, The Bad and The Ugly

Peter Dayan[a] and Yael Niv[b]

Reinforcement learning provides both qualitative and quantitative frameworks for understanding and modeling adaptive decision-making in the face of rewards and punishments. Here we review the latest dispatches from the forefront of this field, and map out some of the territories where lie monsters.

Addresses
[a] UCL, United Kingdom
[b] Psychology Department and Princeton Neuroscience Institute, Princeton University, United States

Corresponding authors: Dayan, Peter (dayan@gatsby.ucl.ac.uk) and Niv, Yael (yael@princeton.edu)

Current Opinion in Neurobiology 2008, **18**:185–196

This review comes from a themed issue on
Cognitive neuroscience
Edited by Read Montague and John Assad

Available online 22nd August 2008

0959-4388/$ – see front matter

DOI 10.1016/j.conb.2008.08.003

Introduction

Reinforcement learning (RL) [1] studies the way that natural and artificial systems can learn to predict the consequences of and optimize their behavior in environments in which actions lead them from one state or situation to the next, and can also lead to rewards and punishments. Such environments arise in a wide range of fields, including ethology, economics, psychology, and control theory. Animals, from the most humble to the most immodest, face a range of such optimization problems [2], and, to an apparently impressive extent, solve them effectively. RL, originally born out of mathematical psychology and operations research, provides qualitative and quantitative computational-level models of these solutions.

However, the reason for this review is the increasing realization that RL may offer more than just a computational, 'approximate ideal learner' theory for affective decision-making. RL algorithms, such as the temporal difference (TD) learning rule [3], appear to be directly instantiated in neural mechanisms, such as the phasic activity of dopamine neurons [4]. That RL appears to be so transparently embedded has made it possible to use it in a much more immediate way to make hypotheses about, and retrodictive and predictive interpretations of, a wealth of behavioral and neural data collected in a huge range of paradigms and systems. The literature in this area is by now extensive, and has been the topic of many recent reviews (including [5–9]). This is in addition to rapidly accumulating literature on the partly related questions of optimal decision-making in situations involving slowly amounting information, or social factors such as games [10–12]. Thus here, after providing a brief sketch of the overall RL scheme for control (for a more extensive review, see [13]), we focus only on some of the many latest results relevant to RL and its neural instantiation. We categorize these recent findings into those that fit comfortably with, or flesh out, accepted notions (playfully, 'The Good'), some new findings that are not as snugly accommodated, but suggest the need for extensions or modifications ('The Bad'), and finally some key areas whose relative neglect by the field is threatening to impede its further progress ('The Ugly').

The reinforcement learning framework

Decision-making environments are characterized by a few key concepts: a state space (states are such things as locations in a maze, the existence or absence of different stimuli in an operant box or board positions in a game), a set of actions (directions of travel, presses on different levers, and moves on a board), and affectively important outcomes (finding cheese, obtaining water, and winning). Actions can move the decision-maker from one state to another (i.e. induce state transitions) and they can produce outcomes. The outcomes are assumed to have numerical (positive or negative) utilities, which can change according to the motivational state of the decision-maker (e.g. food is less valuable to a satiated animal) or direct experimental manipulation (e.g. poisoning). Typically, the decision-maker starts off not knowing the rules of the environment (the transitions and outcomes engendered by the actions), and has to learn or sample these from experience.

In instrumental conditioning, animals learn to choose actions to obtain rewards and avoid punishments, or, more generally to achieve goals. Various goals are possible, such as optimizing the average rate of acquisition of net rewards (i.e. rewards minus punishments), or some proxy for this such as the expected sum of future rewards, where outcomes received in the far future are *discounted* compared with outcomes received more immediately. It is the long-term nature of these goals that necessitates consideration not only of immediate outcomes but also of state transitions, and makes choice interesting and difficult. In terms of RL, instrumental conditioning concerns optimal choice, that is determining an assignment of actions to states, also known as a *policy* that optimizes the subject's goals.

By contrast with instrumental conditioning, Pavlovian (or classical) conditioning traditionally treats how subjects learn to predict their fate in those cases in which they cannot actually influence it. Indeed, although RL is primarily concerned with situations in which action selection is germane, such predictions play a major role in assessing the effects of different actions, and thereby in optimizing policies. In Pavlovian conditioning, the predictions also lead to responses. However, unlike the flexible policies that are learned via instrumental conditioning, Pavlovian responses are hard-wired to the nature and emotional valence of the outcomes.

RL methods can be divided into two broad classes, *model-based* and *model-free*, which perform optimization in very different ways (Box 1, [14]). Model-based RL uses experience to construct an internal model of the transitions and immediate outcomes in the environment. Appropriate actions are then chosen by searching or planning in this world model. This is a statistically efficient way to use experience, as each morsel of information from the environment can be stored in a statistically faithful and computationally manipulable way. Provided that constant replanning is possible, this allows action selection to be readily adaptive to changes in the transition contingencies and the utilities of the outcomes. This flexibility makes model-based RL suitable for supporting goal-directed actions, in the terms of Dickinson and Balleine [15]. For instance, in model-based RL, performance of actions leading to rewards whose utilities have decreased is immediately diminished. Via this identification and other findings, the behavioral neuroscience of such goal-directed actions suggests a key role in model-based RL (or at least in its components such as outcome evaluation) for the dorsomedial striatum (or its primate homologue, the caudate nucleus), prelimbic prefrontal cortex, the orbitofrontal cortex, the medial prefrontal cortex,[1] and parts of the amygdala [9,17–20].

Model-free RL, on the other hand, uses experience to learn directly one or both of two simpler quantities (state/action values or policies) which can achieve the same optimal behavior but without estimation or use of a world model. Given a policy, a state has a *value*, defined in terms of the future utility that is expected to accrue starting from that state. Crucially, correct values satisfy a set of mutual consistency conditions: a state can have a high value only if the actions specified by the policy at that state lead to immediate outcomes with high utilities, and/or states which promise large future expected utilities (i.e. have high values themselves). Model-free learning rules such as the temporal difference (TD) rule [3] define any momentary *inconsistency* as a prediction error, and use it to specify rules for plasticity that allow learning of more accurate values and decreased inconsistencies. Given correct values, it is possible to improve the policy by preferring those actions that lead to higher utility outcomes and higher valued states. Direct model-free methods for improving policies without even acquiring values are also known [21].

Model-free methods are statistically less efficient than model-based methods, because information from the environment is combined with previous, and possibly erroneous, estimates or beliefs about state values, rather than being used directly. The information is also stored in scalar quantities from which specific knowledge about rewards or transitions cannot later be disentangled. As a result, these methods cannot adapt appropriately quickly to changes in contingency and outcome utilities. Based on the latter characteristic, model-free RL has been suggested as a model of habitual actions [14,15], in which areas such as the dorsolateral striatum and the amygdala are believed to play a key role [17,18]. However, a far more direct link between model-free RL and the workings of affective decision-making is apparent in the findings that the phasic activity of dopamine neurons during appetitive conditioning (and indeed the fMRI BOLD signal in the ventral striatum of humans, a key target of dopamine projections) has many of the quantitative characteristics of the TD prediction error that is key to learning model-free values [4,22–24]. It is these latter results that underpin the bulk of neural RL.

'The Good': new findings in neural RL

Daw *et al.* [5] sketched a framework very similar to this, and reviewed the then current literature which pertained to it. Our first goal is to update this analysis of the literature. In particular, courtesy of a wealth of experiments, just two years later we now know much more about the functional organization of RL systems in the brain, the pathways influencing the computation of prediction errors, and time-discounting. A substantial fraction of this work involves mapping the extensive findings from rodents and primates onto the human brain, largely using innovative experimental designs while measuring the fMRI BOLD signal. This research has proven most fruitful, especially in terms of tracking prediction error signals in the human brain. However, in considering these results it is important to remember that the BOLD signal is a measure of oxyhemoglobin levels and not neural activity, let alone dopamine. That neuromodulators can act directly on capillary dilation, and that even the limited evidence we have about the coupling between synaptic drive or neural activity and the BOLD signal is confined to the cortex (e.g. [25]) rather than the striatum or mid-

[1] Note that here and henceforth we lump together results from rat and primate prefrontal cortical areas for the sake of brevity and simplicity, despite important and contentious differences [16].

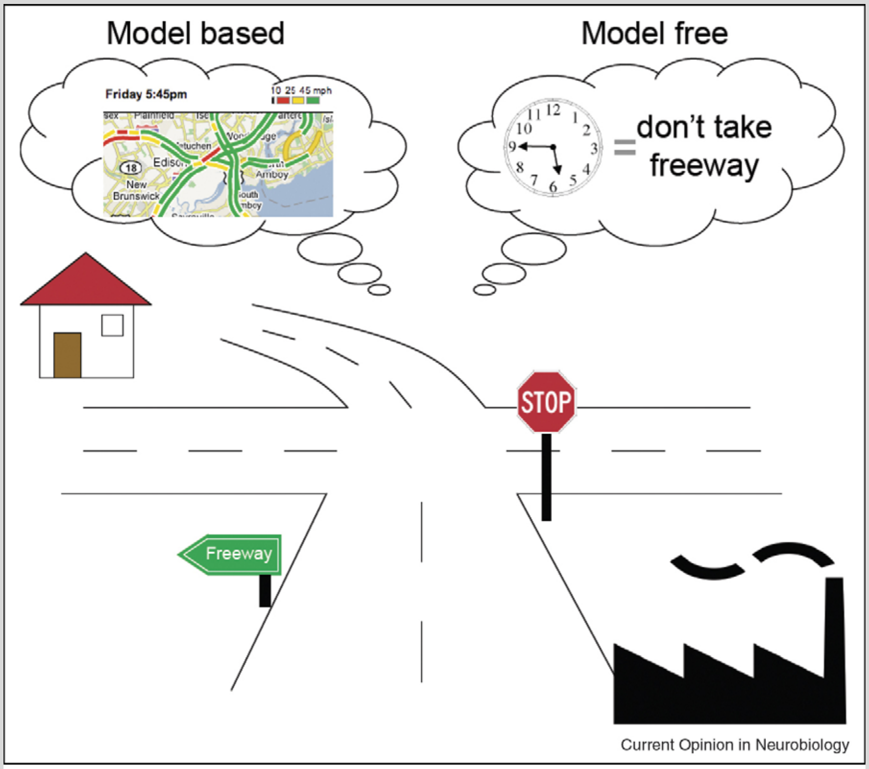
Box 1 Model-based and model-free reinforcement learning

Reinforcement learning methods can broadly be divided into two classes, model-based and model-free. Consider the problem illustrated in the figure, of deciding which route to take on the way home from work on Friday evening. We can abstract this task as having states (in this case, locations, notably of junctions), actions (e.g. going straight on or turning left or right at every intersection), probabilities of transitioning from one state to another when a certain action is taken (these transitions are not necessarily deterministic, e.g. due to road works and bypasses), and positive or negative outcomes (i.e. rewards or costs) at each transition from scenery, traffic jams, fuel consumed, etc. (which are again probabilistic).

Model-based computation, illustrated in the left 'thought bubble', is akin to searching a mental map (a *forward model* of the task) that has been learned based on previous experience. This forward model comprises knowledge of the characteristics of the task, notably, the probabilities of different transitions and different immediate outcomes. Model-based action selection proceeds by searching the mental map to work out the long-run value of each action at the current state in terms of the expected reward of the whole route home, and chooses the action that has the highest value.

Model-free action selection, by contrast, is based on learning these long-run values of actions (or a preference order between actions) without either building or searching through a model. RL provides a number of methods for doing this, in which learning is based on momentary inconsistencies between successive estimates of these values along sample trajectories. These values, sometimes called *cached* values because of the way they store experience, encompass all future probabilistic transitions and rewards in a single scalar number that denotes the overall future worth of an action (or its attractiveness compared with other actions). For instance, as illustrated in the right 'thought bubble', experience may have taught the commuter that on Friday evenings the best action at this intersection is to continue straight and avoid the freeway.

Model-free methods are clearly easier to use in terms of online decision-making; however, much trial-and-error experience is required to make the values be good estimates of future consequences. Moreover, the cached values are inherently inflexible: although hearing about an unexpected traffic jam on the radio can immediately affect action selection that is based on a forward model, the effect of the traffic jam on a cached propensity such as 'avoid the freeway on Friday evening' cannot be calculated without further trial-and-error learning on days in which this traffic jam occurs. Changes in the goal of behavior, as when moving to a new house, also expose the differences between the methods: whereas model-based decision making can be immediately sensitive to such a goal-shift, cached values are again slow to change appropriately. Indeed, many of us have experienced this directly in daily life after moving house. We clearly know the location of our new home, and can make our way to it by concentrating on the new route; but we can occasionally take an habitual wrong turn toward the old address if our minds wander. Such introspection, and a wealth of rigorous behavioral studies (see [15], for a review) suggests that the brain employs both model-free and model-based decision-making strategies in parallel, with each dominating in different circumstances [14]. Indeed, somewhat different neural substrates underlie each one [17].



Current Opinion in Neurobiology

Figure 1: Two ways to choose which route to take when traveling home from work on friday evening.

brain areas such as the ventral tegmental area, imply that this fMRI evidence is alas very indirect.

Functional organization: in terms of the mapping from rodents to humans, reinforcer devaluation has been employed to study genuinely goal-directed choice in humans, that is, to search for the underpinnings of behavior that is flexibly adjusted to such things as changes in the value of a predicted outcome. The orbitofrontal cortex (OFC) has duly been revealed as playing a particular role in representing goal-directed value [26•]. However, the bulk of experiments has not set out to distinguish model-based from model-free systems, and has rather more readily found regions implicated in model-free processes. There is some indirect evidence [27–29,30•] for the involvement of dopamine and dopaminergic mechanisms in learning from reward prediction errors in humans, along with more direct evidence from an experiment involving the pharmacological manipulation of dopamine [31•]. These studies, in which prediction errors have been explicitly modeled, along with others, which use a more general multivariate approach [32] or an argument based on a theoretical analysis of multiple, separate, cortico-basal ganglia loops [33], also reveal roles for OFC, medial prefrontal cortical structures, and even the cingulate cortex. The contributions of the latter especially have recently been under scrutiny in animal studies focused on the cost-benefit tradeoffs inherent in decision-making [34,35], and the involvement of dopaminergic projections to and from the anterior cingulate cortex, and thus potential interactions with RL have been suggested ([36], but see [37]). However, novel approaches to distinguishing model-based and model-free control may be necessary to tease apart more precisely the singular contributions of the areas.

Computation of prediction errors: in terms of pathways, one of the more striking recent findings is evidence that the lateral habenula suppresses the activity of dopamine neurons [38•,39], in a way which may be crucial for the representation of the negative prediction errors that arise when states turn out to be worse than expected. One natural possibility is then that pauses in the burst firing of dopamine cells might code for these negative prediction errors. This has received some quantitative support [40], despite the low baseline rate of firing of these neurons, which suggests that such a signal would have a rather low bandwidth. Further, studies examining the relationship between learning from negative and positive consequences and genetic polymorphisms related to the D2 dopaminergic receptor have established a functional link between dopamine and learning when outcomes are worse than expected [41,42]. Unfortunately, marring this apparent convergence of evidence is the fact that the distinction between the absence of an expected reward and more directly aversive events is still far from being clear, as we complain below.

Further data on contributions to the computation of the TD prediction error have come from new findings on excitatory pathways into the dopamine system [43]. Evidence about the way that the amygdala [44,45•,46] and the medial prefrontal cortex [47] code for both positive and negative predicted values and errors, and the joint coding of actions, the values of actions, and rewards in striatal and OFC activity [48–55,9], is also significant, given the putative roles of these areas in learning and representing various forms of RL values.

In fact, the machine learning literature has proposed various subtly different versions of the TD learning signal, associated with slightly different model-free RL methods. Recent evidence from a primate study [56•] looking primarily at one dopaminergic nucleus, the substantia nigra pars compacta, seems to support a version called SARSA [57]; whereas evidence from a rodent study [58•] of the other major dopaminergic nucleus, the ventral tegmental area, favors a different version called *Q*-learning (CJCH Watkins, Learning from delayed rewards, PhD thesis, University of Cambridge, 1989). Resolving this discrepancy, and indeed, determining whether these learning rules can be incorporated within the popular Actor/Critic framework for model-free RL in the basal ganglia [59,1], will necessitate further experiments and computational investigation.

A more radical change to the rule governing the activity of dopamine cells which separates out differently the portions associated with the outcome (the primary reward) and the learned predictions has also been suggested in a modeling study [60]. However, various attractive features of the TD rule, such as its natural account of secondary conditioning and its resulting suitability for optimizing sequences of actions leading up to a reward, are not inherited directly by this rule.

Temporal discounting: a recurrent controversy involves the way that the utilities of proximal and distant outcomes are weighed against each other. Exponential discounting, similar to a uniform interest rate, has attractive theoretical properties, notably the absence of intertemporal choice conflict, the possibility of recursive calculation scheme and simple prediction errors [1]. However, the more computationally complex hyperbolic discounting, which shows preference reversals and impulsivity, is a more common psychological finding in humans and animals [61].

The immediate debate concerned the abstraction and simplification of hyperbolic discounting to two evaluative systems, one interested mostly in the here and now, the other in the distant future [62], and their apparent instantiation in different subcortical and cortical structures respectively [63,64]. This idea became somewhat wrapped-up with the distinct notion that these neural areas are involved in model-free and model-based learning. Further, other studies found a more unitary neural representation of discounting [65], at least in the BOLD signal, and recent results confirm that dopaminergic prediction errors indeed show the expected effects of discounting [58•]. One surprising finding is that the OFC may separate out the representation of the temporal discount factor applied to distant rewards from that of the magnitude of the reward [54], implying a complex problem of how these quantities are then integrated.

There has also been new work based on the theory [8] that the effective interest rate for time is under the influence of the neuromodulator serotonin. In a task that provides a fine-scale view of temporal choice [66], dietary reduction of serotonin levels in humans (tryptophan depletion) gave rise to extra impulsivity, that is, favoring smaller rewards sooner over larger rewards later, which can be effectively modeled by a steeper interest rate [67]. Somewhat more problematically for the general picture of control sketched above, tryptophan depletion also led to a topographically mapped effect in the striatum, with quantities associated with predictions for high interest rates preferentially correlated with more ventral areas, and those for low interest rates with more dorsal areas [68].

The idea that subjects are trying to optimize their long-run rates of acquisition of reward has become important in studies of time-sensitive sensory decision-making [11,69]. It also inspired a new set of modeling investigations into free operant choice tasks, in which subjects are free to execute actions at times of their own choosing, and the dependent variables are quantities such as the rates of responding [70,71]. In these accounts, the rate of reward acts as an opportunity cost for time, thereby penalizing sloth, and is suggested as being coded by the tonic (as distinct from the phasic) levels of dopamine. This captures findings associated with response vigor given dopaminergic manipulations [72,73,74•], and, at least assuming a particular coupling between phasic and tonic dopamine, can explain results linking vigor to predictions of reward [75,76].

'The Bad': apparent but tractable inconsistencies

Various research areas which come in close contact with different aspects of RL, help extend or illuminate it in not altogether expected ways. These include issues of aversive-appetitive interactions, exploration and novelty, a range of phenomena important in neuroeconomics [77,78] such as risk, Pavlovian-instrumental interactions, and also certain new structural or architectural findings. The existence of multiple control mechanisms makes it challenging to interpret some of these results unambiguously, since rather little is known for sure about the interaction between model-free and model-based systems, or the way the neural areas involved communicate, cooperate and compete.

Appetitive–aversive interactions: one key issue that dogs neural RL [79] is the coding of aversive rather than appetitive prediction errors. Although dopamine neurons are seemingly mostly inhibited by unpredicted punishments [80,81], fMRI studies into the ventral striatum in humans have produced mixed results, with aversive prediction errors sometimes leading to above-baseline BOLD [82,83], but other times below-baseline BOLD, perhaps with complex temporal dynamics [84,23]. Dopamine antagonists suppress the aversive prediction error signal [85], and withdrawing (OFF) or administering (ON) dopamine-boosting medication to patients suffering from Parkinson's disease leads to boosted and suppressed learning from negative outcomes, respectively [86].

One study that set out to compare directly appetitive and aversive prediction errors found a modest spatial separation in the ventral striatal BOLD signal [87], consistent with various other findings about the topography of this structure [88,89]; indeed, there is a similar finding in the OFC [90]. However, given that nearby neurons in the amygdala code for either appetitive or aversive outcomes [44,45•,46], fMRI's spotlight may be too coarse to address all such questions adequately.

Aversive predictions are perhaps more complex than appetitive ones, owing to their multifaceted range of effects, with different forms of contextual information (such as defensive distance; [91]) influencing the choice between withdrawal and freezing versus approach and fighting. Like appetitive choice behavior, some of these behavioral effects require vigorous actions, which we discussed above in terms of tonic levels of dopamine [70,71]. Moreover, aversive predictions can lead to active avoidance, which then brings about appetitive predictions associated with the achievement of safety. This latter mechanism is implicated in conditioned avoidance responses [92], and has also been shown to cause increases in BOLD in the same part of the OFC that is also activated by the receipt of rewards [93].

Novelty, uncertainty and exploration: the bulk of work in RL focuses on exploitation, that is on using past experience to optimize outcomes on the next trial or next period. More ambitious agents seek also to optimize exploration, taking account in their choices not only the benefits of known (expected) future rewards, but also the potential benefits of learning about unknown rewards and punishments (i.e. the long-term gain to be harvested due to acquiring knowledge about the values of different states). The balancing act between these requires careful accounting for uncertainty, in which the neuromodulators acetylcholine and norepinephrine have been implicated

[94,95]. Uncertainty is also related to novelty, which seems to involve dopaminergic areas and some of their targets [96•,97,98]. Uncertainty and novelty can support different types of exploration, respectively, directed exploration, in which exploratory actions are taken in proportion to known uncertainty, and undirected exploration, in which novel unknown parts of the environment are uniformly explored. Tracking uncertainty has been shown to involve the PFC in humans [99] and complex patterns of activity in lateral and medial prefrontal cortex of monkeys [100], and has also been associated with functions of the amygdala [101].

Model-free and model-based systems in RL adopt rather different approaches to exploration, although because uncertainty generally favors model-based control [14], the model-free approaches may be concealed. Indeed, we may expect that certain forms of explicit change and uncertainty could act to transfer habitized control back to a model-based system [102]. Model-based exploration probably involves frontal areas, as exemplified by one recent imaging study which revealed a particular role for fronto-polar cortex in trials in which nonexploitative actions were chosen [103]. Model-free exploration has been suggested to involve neuromodulatory systems such as dopamine and norepinephrine, instantiating computationally more primitive strategies such as dopaminergic exploration bonuses [98].

Uncertainty should influence learning as well as exploration. In particular, learning (and forgetting) rates should be higher in a rapidly changing environment and slower in a rather stationary one. A recent fMRI study demonstrated that human subjects adjust their learning rates according to the volatility of the environment, and suggested that the anterior cingulate cortex is involved in the online tracking of volatility [104]. In animal conditioning tasks involving a more discrete form of surprise associated with stimuli, substantial studies have shown a role for the central nucleus of the amygdala and the cholinergic neurons in the substantia innominata and nucleus basalis in upregulating subsequent learning associated with that stimulus [94]. More recent work has shown that there is a crucial role for the projection from the dopaminergic (putatively prediction error coding) substantia nigra pars compacta to the central nucleus in this upregulation [105], and that the involvement of the central nucleus is limited to the time of surprise (and that of the cholinergic nuclei to the time of subsequent learning; [106]). This accords with evidence that some amygdala neurons respond directly to both positive and negative surprising events ([45•], a minimum requirement for a surprise signal) and to unpredictability itself [101].

Risk, regret and neuroeconomics: the relatively straightforward view of utility optimization that permeates neural RL is challenged by various findings in and around the fields of experimental economics, behavioral economics, and neuroeconomics [107,77]. These study the psychological and neural factors underlying a range of situations in which behavior departs from apparent normative ideals. One example of this concerns the sensitivity of choice behavior to *risk* associated with variability in the outcomes. There are extensive economic theories about risk, and indeed evidence from functional neuroimaging that variability may be coded in specific cortical and subcortical regions such as the insular cortex, OFC and the ventral striatum [108–112]. However, how risk should influence action selection in model-free and model-based control is not completely clear, partly because variability can have both an indirect influence on values, for instance via a saturating utility function for rewards [113], and direct effects through a risk-sensitive learning process [114,115]. Although nonlinear utilities for rewards will affect both types of controllers, the effects of risk through sampling biases [114] or through a learning rule that is asymmetrically sensitive to positive and negative errors [115] depend on the specific implementation of learning in a model-free or model-based controller, and thus can differ between the two.

A second neuroeconomic concern is *regret*, in which a form of counterfactual thinking [116] or fictive learning [117] is induced when foregone alternatives turn out to be better than those that were chosen [118,119]. Imaging studies have suggested a role for the OFC and other targets of the dopamine system in processing fictive learning signals [120,121,117]. Here, also, there may be separate instantiations in model-based and model-free systems, as accounting for counterfactuals is straightforward in the former, but requires extra machinery in the latter.

A third issue along these lines is *framing*, in which different descriptions of a single (typically risky) outcome result in different valuations. Subjects are more reluctant to take chances in the face of apparent gains than losses, even when the options are actually exactly the same. A recent imaging study [122] reported a close correlation between susceptibility to framing and activity in the amygdala, and a negative correlation with activity in OFC and medial PFC.

Finally, there is an accumulating wealth of work on social utility functions, and game–theoretic interactions between subjects [123], including studies in patient populations with problems with social interactions [124]. There are interesting hints for neural RL from studies that use techniques such as transcranial magnetic stimulation to disrupt specific prefrontal activity, which turns out to affect particular forms of social choice such as the ability to reject patently unfair (though still lucrative) offers in games of economic exchange [125]. However, these have yet to be coupled to the nascent more complete RL accounts of the complex behavior in pair and multi-agent tournaments [126,127].

Pavlovian values: although it is conventional in RL to consider Pavlovian or classical conditioning as only being about the acquisition of predictions, with instrumental conditioning being wholly responsible for choice, it is only through a set of (evolutionarily programmed) responses, such as the approach that is engendered by predictions of reward, that Pavlovian predictions become evident in the first place. Indeed, Pavlovian responses can famously out-compete instrumental responses [128]. This is of particular importance in omission schedules or negative automaintenance, when subjects continue responding based on predictions of future rewards, even when their actions actually prevent them from getting those rewards.

Other paradigms such as Pavlovian-instrumental transfer, in which Pavlovian predictors influence the vigor of instrumental responding, are further evidence of non-normative interactions among these two forms of learning, potentially accounting for some of the effects of manipulating reward schedules on effort [52,51]. Part of the architecture of this transfer involving the amygdala, the striatum and dopamine that has been worked out in rodent studies [17] has recently been confirmed in a human imaging experiment [129]; however, a recent surprise was that lesions in rats that disconnected the central nucleus of the amygdala from the ventral tegmental area actually enhanced transfer [130], rather than suppressing it, as would have been expected from earlier work (e.g. [131]).

The Pavlovian preparatory response to aversive predictions is often withdrawal or disengagement. It has been hypothesized that this is a form of serotonergically mediated inhibition [132,133] that acts to arrest paths leading to negative outcomes, and, in model-based terms, excise potentially aversive parts of the search tree [134]. Compromising serotonin should thus damage this reflexive inhibition, leading to systematic problems with evaluation and choice. The consummatory Pavlovian response to immediate threat is panic, probably mediated by the peri-acqueductal gray [135,91]; such responses have themselves recently been observed in an imaging study which involved a virtual threatening 'chase' [136].

Appetitive and aversive Pavlovian influences have been suggested as being quite pervasive [137], accounting for aspects of impulsivity apparent in hyperbolic discounting, framing and the like. Unfortunately, the sophisticated behavioral and neural analyses of model-free and model-based instrumental values are not paralleled, as of yet, by an equivalently worked-out theory for the construction of Pavlovian values. Moreover, Pavlovian influences may affect model-based and model-free instrumental actions differently, suggesting even more relatively untapped complexity.

Structural findings: finally, there are some recent fascinating and sometimes contrarian studies into the way that striatal pathways function in choice. One set [138•,139,140] has revealed a role for the subthalamic nucleus (STN) in slowing down choices between strongly appetitive options, perhaps to give more time for the precise discrimination of the best action from merely good ones [141]. It could again be that the impulsivity created when the STN is suppressed arises from a form of Pavlovian approach [137]. Others have looked at serial processes that shift action competition and choice up a ventral to dorsal axis along the striatum [142,143], perhaps consistent with the suggested spiraling connectivity through dopaminergic nuclei [144,145].

'The Ugly': crucial challenges

The last set of areas of neural RL suffer a crucial disparity between their importance and the relative dearth of systematic or comprehensive studies. Hopefully, by the next review, this imbalance will be at least partially redressed.

In terms of model-based control, a central question concerns the acquisition and use of hierarchical structures. It seems rather hopeless to plan using only the smallest units of action (think of the twitches of a single muscle), and so there is a great interest in hierarchies in both standard [146] and neural RL [147]. Unfortunately, the crucial problem of acquiring appropriate hierarchical structure in the absence of supervision is far from being solved. A second concern has to do with a wide class of learning scenarios in which optimal learning relies on detection of change and, essentially, learning of a 'new' situation, rather than updating the previously learned one. A prime example of this is the paradigm of extinction in which a cue previously predictive of reward is no longer paired with the rewarding outcome. Behavioral results show that animals and humans do not simply 'unlearn' the previous predictive information, but rather they learn a new predictive relationship that is inhibitory to the old one [148]. How this type of learning is realized within a RL framework, whether model-based or model-free, is at present unclear, though some suggestions have been made [149]. The issues mentioned above to do with uncertainty and change also bear on this.

In terms of model-free control, there are various pressing anomalies (on top of appetitive/aversive opponency) associated with the activity of dopamine cells (and the not-necessarily equivalent release of dopamine at its targets; [150]). One concerns prediction errors inspired by conditioned inhibitors for reward and punishment predictors. While the predictive stimuli we have contemplated so far are stimuli that predict the occurrence of some outcome, a conditioned inhibitor is a stimulus that predicts that an otherwise expected outcome will *not* happen. How the former (conditioned excitors) and the latter (conditioned inhibitors) interact to generate a single prediction and a corresponding prediction error is not yet clear either neurally or computationally. A fascinating electrophysiological study [151] targeting this issue suggests that the answer is not simple.

A second anomaly is the apparent adaptive scaling of prediction errors depending on the precise range of such errors that is expected, forming an informationally efficient code [152]. As with other cases of adaptation, it is not clear that downstream mechanisms (here, presumably dopamine receptors) have the information or indeed the calculational wherewithal to reverse engineer correctly the state of adaptation of the processes controlling the activity of dopamine cells. This opens up the possibility that there might be biases in the interpretation of the signal, leading to biases in choice. Alternatively, this adaptation might serve a different computational goal, such as adjusting learning appropriately in the face of forms of uncertainty [153].

A timely reminder of the complexity of the dopamine system came from a recent study suggesting a new subclass of dopamine neurons with particularly unusual neurophysiological properties, that selectively target the prefrontal cortex, the basolateral amygdala and the core of the accumbens [154]. It is important to remember that most of our knowledge about the behaviorally relevant activity of dopamine neurons comes from recordings during Pavlovian tasks or rather simple instrumental scenarios — basic questions about dopaminergic firing patterns when several actions are necessary in order to obtain an outcome, when there are several outcomes in a trial, or in hierarchical tasks are still unanswered. Lammel *et al.*'s [154] results add to previous subtle challenges to the common premise that the dopaminergic signal is unitary (e.g. when taking together [56•,58•]). Future physiological studies and recordings in complex tasks are needed to flesh out what could potentially be a more diverse signal than has so far been considered. The role of dopamine in the prefrontal cortex, which has only played prominently in relatively few models (notably [155]), and indeed its potential effects on the model-based controller, might also need to be rethought in light of these findings.

A final area of unfortunate lack of theory and dearth of data has to do with timing. The activity pattern of dopamine cells that most convincingly suggests that they report a prediction error is the well-timed dip in responding when an expected reward is not provided [156,40]. The neural mechanisms underlying this timing are most unclear; in fact there are various quite different classes of suggestion in the RL literature and beyond. What is worse is that this form of interval timing is subject to substantial scalar noise [157], to a degree that could make accurate prediction learning, and accurately timed dips in dopamine responding, quite challenging to achieve.

Conclusions

As should be apparent from this review, neural RL is a vibrant and dynamic field, generating new results at a near-overwhelming rate, and spreading its wings well beyond its initial narrow confines of trial-and-error reward learning. We have highlighted many foci of ongoing study, and also some orphaned areas mentioned in the previous section. However, our best hope is that the sterling efforts to link together the substantial theoretically motivated and informative animal studies to human neuroimaging results, along with new data from cyclic voltammetric measurements of phasic dopamine concentrations [158], imminent results on serotonin, and even nascent efforts to activate DA cells *in vivo* using new optogenetic methods such as targetted channel rhodopsin, will start to pay off in the opposite direction by deciding between, and forcing improvements to, RL models.

Acknowledgements

We are very grateful to Peter Bossaerts, Nathaniel Daw, Michael Frank, Russell Poldrack, Daniel Salzman, Ben Seymour and Wako Yoshida for their helpful comments on a previous version of this manuscript. PD is funded by the Gatsby Charitable Foundation; YN is funded by the Human Frontiers Science Program.

References and recommended reading

Papers of particular interest, published within the period of the review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Sutton RS, Barto AG: **Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)**The MIT Press; 1998.

2. Montague PR: **Why Choose This Book?: How We Make Decisions**.Dutton Adult; 2006.

3. Sutton R: **Learning to predict by the methods of temporal differences**. *Mach Learn* 1988, **3**:9-44.

4. Montague PR, Dayan P, Sejnowski TJ: **A framework for mesencephalic dopamine systems based on predictive Hebbian learning**. *J Neurosci* 1996, **16**:1936-1947.

5. Daw ND, Doya K: **The computational neurobiology of learning and reward**. *Curr Opin Neurobiol* 2006, **16**:199-204.

6. Johnson A, van der Meer MA, Redish AD: **Integrating hippocampus and striatum in decision-making**. *Curr Opin Neurobiol* 2007, **17**:692-697.

7. 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.

8. Doya K: **Modulators of decision making**. *Nat Neurosci* 2008, **11**:410-416.

9. Rushworth MFS, Behrens TEJ: **Choice, uncertainty and value in prefrontal and cingulate cortex**. *Nat Neurosci* 2008, **11**:389-397.

10. Körding K: **Decision theory: what "should" the nervous system do?** *Science* 2007, **318**:606-610.

11. Gold JI, Shadlen MN: **The neural basis of decision making**. *Annu Rev Neurosci* 2007, **30**:535-574.

12. Lee D: **Neural basis of quasi-rational decision making**. *Curr Opin Neurobiol* 2006, **16**:191-198.

13. Niv Y, Montague PR: **Theoretical and empirical studies of learning**. In *Neuroeconomics: Decision Making and The Brain*. Edited by Glimcher PW, Camerer C, Fehr E, Poldrack R. New York, NY: Academic Press; 2008:329–349.

14. Daw ND, Niv Y, Dayan P: **Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control**. *Nat Neurosci* 2005, **8**:1704-1711.

15. Dickinson A, Balleine B: **The role of learning in motivation**. In *Stevens' Handbook of Experimental Psychology*. Edited by Gallistel C. New York, NY: Wiley; 2002:497–533.

16. Uylings HBM, Groenewegen HJ, Kolb B: **Do rats have a prefrontal cortex?** *Behav Brain Res* 2003, **146**:3-17.

17. Balleine BW: **Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits**. *Physiol Behav* 2005, **86**:717-730.

18. Killcross S, Coutureau E: **Coordination of actions and habits in the medial prefrontal cortex of rats**. *Cereb Cortex* 2003, **13**:400-408.

19. Dolan RJ: **The human amygdala and orbital prefrontal cortex in behavioural regulation**. *Philos Trans R Soc Lond B: Biol Sci* 2007, **362**:787-799.

20. Matsumoto K, Tanaka K: **The role of the medial prefrontal cortex in achieving goals**. *Curr Opin Neurobiol* 2004, **14**:178-185.

21. Baxter J, Bartlett P: **Infinite-horizon policy-gradient estimation**. *J Artif Intell Res* 2001, **15**:319-350.

22. Berns GS, McClure SM, Pagnoni G, Montague PR: **Predictability modulates human brain response to reward**. *J Neurosci* 2001, **21**:2793-2798.

23. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ: **Temporal difference models and reward-related learning in the human brain**. *Neuron* 2003, **38**:329-337.

24. Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, Samejima K, Imamizu H, Kawato M: **A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task**. *J Neurosci* 2004, **24**:1660-1665.

25. Logothetis NK, Wandell BA: **Interpreting the BOLD signal**. *Annu Rev Physiol* 2004, **66**:735-769.

26. • Valentin VV, Dickinson A, O'Doherty JP: **Determining the neural substrates of goal-directed learning in the human brain**. *J Neurosci* 2007, **27**:4019-4026. Rodent behavioral paradigms [15] have played a crucial role in establishing the structural and functional dissociations between different classes of behavior [14]. This paper is one of a series of studies from this group which is importing these paradigms lock, stock and barrel to humans, showing that exactly the same distinctions and indeed homologous anatomical bases apply. These studies form a crucial part of the evidentiary basis of neural RL.

27. O'Doherty JP, Buchanan TW, Seymour B, Dolan RJ: **Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum**. *Neuron* 2006, **49**:157-166.

28. 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**:12860-12867.

29. Tobler PN, O'Doherty JP, Dolan RJ, Schultz W: **Human neural learning depends on reward prediction errors in the blocking paradigm**. *J Neurophysiol* 2006, **95**:301-310.

30. • D'Ardenne K, McClure SM, Nystrom LE, Cohen JD: **Bold responses reflecting dopaminergic signals in the human ventral tegmental area**. *Science* 2008, **319**:1264-1267. This study achieves an impressive new level of anatomical precision in recording fMRI BOLD signals from the ventral tegmental area (VTA) in humans, confirming various aspects suggested from recordings in macaques and rats (though the extra precision does not extend to a distinction between dopaminergic and nondopaminergic cells in the VTA, let alone the different classes of dopamine cells). It will be fascinating to see what happens in the substantia nigra pars compacta, whose dopamine cells may be more closely associated with action rather than value learning..

31. • Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD: **Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans**. *Nature* 2006, **442**:1042-1045. Despite the substantial circumstantial data from animals and human neurological cases, there are few direct tests of the specifically dopaminergic basis of normal instrumental conditioning. This study tested this in healthy subjects using L-DOPA and haloperidol, which are expected respectively to enhance and suppress dopamine function. L-DOPA improved choice performance toward monetary gains (compared to haloperidol-treated subjects) but not avoidance of monetary losses. This was accompanied by an increase in ventral striatal BOLD responses corresponding to prediction errors in the gain condition in the L-DOPA group, whose magnitude was sufficient for a standard action-value learning model to explain the effects of the drugs on behavioral choices. These results provide (much needed, but still indirect) evidence that the ventral striatal prediction error BOLD signal indeed reflects (at least partly) dopaminergic activity.

32. Hampton AN, O'Doherty JP: **Decoding the neural substrates of reward-related decision making with functional MRI**. *Proc Natl Acad Sci U S A* 2007, **104**:1377-1382.

33. Samejima K, Doya K: **Multiple representations of belief states and action values in corticobasal ganglia loops**. *Ann NY Acad Sci* 2007, **1104**:213-228.

34. Walton ME, Bannerman DM, Alterescu K, Rushworth MFS: **Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions**. *J Neurosci* 2003, **23**:6475-6479.

35. Schweimer J, Hauber W: **Involvement of the rat anterior cingulate cortex in control of instrumental responses guided by reward expectancy**. *Learn Mem* 2005, **12**:334-342.

36. Schweimer J, Hauber W: **Dopamine D1 receptors in the anterior cingulate cortex regulate effort-based decision making**. *Learn Mem* 2006, **13**:777-782.

37. Walton ME, Croxson PL, Rushworth MFS, Bannerman DM: **The mesocortical dopamine projection to anterior cingulate cortex plays no role in guiding effort-related decisions**. *Behavioral Neuroscience* 2005, **119**:323-328.

38. • Matsumoto M, Hikosaka O: **Lateral habenula as a source of negative reward signals in dopamine neurons**. *Nature* 2007, **447**:1111-1115. The sources of excitation and inhibition of dopaminergic and indeed nondopaminergic cells in the VTA have long been debated. This paper shows that there is an important pathway from a nucleus called the lateral habenula, that has the effect of inhibiting the activity of the dopamine cells. Consistent with this connection, habenula neurons were excited by stimuli that effectively predict the absence of reward, and inhibited by targets that predict its presence.

39. Lecourtier L, Defrancesco A, Moghaddam B: **Differential tonic influence of lateral habenula on prefrontal cortex and nucleus accumbens dopamine release**. *European Journal Neuroscience* 2008, **27**:1755-1762.

40. Bayer HM, Lau B, Glimcher PW: **Statistics of midbrain dopamine neuron spike trains in the awake primate**. *J Neurophysiol* 2007, **98**:1428-1439.

41. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE: **Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning**. *Proc Natl Acad Sci U S A* 2007, **104**:16311-16316.

42. Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, Ullsperger M: **Genetically determined differences in learning from errors**. *Science* 2007, **318**:1642-1645.

43. McHaffie JG, Jiang H, May PJ, Coizet V, Overton PG, Stein BE, Redgrave P: **A direct projection from superior colliculus to substantia nigra pars compacta in the cat**. *Neuroscience* 2006, **138**:221-234.

44. Paton JJ, Belova MA, Morrison SE, Salzman CD: **The primate amygdala represents the positive and negative value of visual stimuli during learning**. *Nature* 2006, **439**:865-870.

45. • Belova MA, Paton JJ, Morrison SE, Salzman CD: **Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdale**. *Neuron* 2007, **55**:970-984 The amygdala has long been suggested to be an important site for appetitive and aversive learning (in which the difference between valences is crucial) and also a form of uncertainty (in which positive and negative valences are treated together). This study and its colleague [44] provides an interesting view into the coding of valenced and unvalenced predictions and prediction errors in the macaque amygdala. Putting these data together with the rodent studies, with their anatomically and functionally precise dissociations [17] is a pressing task..

46. Salzman CD, Paton JJ, Belova MA, Morrison SE: **Flexible neural representations of value in the primate brain**. *Ann N Y Acad Sci* 2007, **1121**:336-354.

47. Matsumoto M, Matsumoto K, Abe H, Tanaka K: **Medial prefrontal cell activity signaling prediction errors of action values**. *Nat Neurosci* 2007, **10**:647-656.

48. Balleine BW, Delgado MR, Hikosaka O: **The role of the dorsal striatum in reward and decision-making**. *J Neurosci* 2007, **27**:8161-8165.

49. Hikosaka O: **Basal ganglia mechanisms of reward-oriented eye movement**. *Ann N Y Acad Sci* 2007, **1104**:229-249.

50. Lau B, Glimcher PW: **Action and outcome encoding in the primate caudate nucleus**. *J Neurosci* 2007, **27**:14502-14514.

51. Hikosaka O, Nakamura K, Nakahara H: **Basal ganglia orient eyes to reward**. *J Neurophysiol* 2006, **95**:567-584.

52. Simmons JM, Ravel S, Shidara M, Richmond BJ: **A comparison of reward-contingent neuronal activity in monkey orbitofrontal cortex and ventral striatum: guiding actions toward rewards**. *Ann NY Acad Sci* 2007, **1121**:376-394.

53. Padoa-Schioppa C: **Orbitofrontal cortex and the computation of economic value**. *Ann N Y Acad Sci* 2007, **1121**:232-253.

54. Roesch MR, Taylor AR, Schoenbaum G: **Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation**. *Neuron* 2006, **51**:509-520.

55. Furuyashiki T, Holland PC, Gallagher M: **Rat orbitofrontal cortex separately encodes response and outcome information during performance of goal-directed behaviour**. *J Neurosci* 2008, **28**:5127-5138.

56. • Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H: **Midbrain dopamine neurons encode decisions for future action**. *Nat Neurosci* 2006, **9**:1057-1063 Most historical recordings from dopamine cells have been in cases in which there is no meaningful choice between different alternatives. This study in macaques, together with Roesch et al's [58•] in rodents looks specifically at choice. The two studies confirm the general precepts of neural RL based on temporal difference (TD) learning, but use rather different methods, and come to different conclusions about which of the variants of TD is at work..

57. Rummery G, Niranjan M: **On-line Q-learning using connectionist systems**, Tech. Rep. Technical Report CUED/F-INFENG/TR 166, Cambridge University Engineering Department, 1994.

58. • Roesch MR, Calu DJ, Schoenbaum G: **Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards**. *Nat Neurosci* 2007, **10**:1615-1624 Along with work from Hyland and colleagues [159,160], this is one of the few studies looking at the activity of dopamine neurons in rodents. Many of its findings are exactly consistent with neural TD, though it comes to a different conclusion about exactly which TD rule is operational from the monkey experiment performed by Morris et al. [56•]..

59. Barto A: **Adaptive critics and the basal ganglia**. In *Models of Information Processing in the Basal Ganglia*. Edited by Houk J, Davis J, Beiser D. Cambridge, MA: MIT Press; 1995:215–232.

60. O'Reilly RC, Frank MJ, Hazy TE, Watz B: **PVLV: the primary value and learned value Pavlovian learning algorithm**. *Behav Neurosci* 2007, **121**:31-49.

61. Ainslie G: **Breakdown of Will**. Cambridge University Press; 2001.

62. Loewenstein G, Prelec D: **Anomalies in intertemporal choice: Evidence and an interpretation**. *The Quarterly Journal of Economics* 1992, **107**:573-597.

63. McClure SM, Laibson DI, Loewenstein G, Cohen JD: **Separate neural systems value immediate and delayed monetary rewards**. *Science* 2004, **306**:503-507.

64. McClure SM, Ericson KM, Laibson DI, Loewenstein G, Cohen JD: **Time discounting for primary rewards**. *J Neurosci* 2007, **27**:5796-5804.

65. Kable JW, Glimcher PW: **The neural correlates of subjective value during intertemporal choice**. *Nat Neurosci* 2007, **10**:1625-1633.

66. Schweighofer N, Shishida K, Han CE, Okamoto Y, Tanaka SC, Yamawaki S, Doya K: **Humans can adopt optimal discounting strategy under real-time constraints**. *PLoS Comput Biol* 2006, **2**:e152.

67. Schweighofer N, Bertin M, Shishida K, Okamoto Y, Tanaka SC, Yamawaki S, Doya K: **Low-serotonin levels increase delayed reward discounting in humans**. *J Neurosci* 2008, **28**:4528-4532.

68. Tanaka SC, Schweighofer N, Asahi S, Shishida K, Okamoto Y, Yamawaki S, Doya K: **Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum**. *PLoS ONE* 2007, **2**:e1333.

69. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD: **The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks**. *Psychol Rev* 2006, **113**:700-765.

70. Niv Y, Joel D, Dayan P: **A normative perspective on motivation**. *Trends Cogn Sci* 2006, **10**:375-381.

71. Niv Y, Daw ND, Joel D, Dayan P: **Tonic dopamine: opportunity costs and the control of response vigor**. *Psychopharmacology (Berl)* 2007, **191**:507-520.

72. Salamone JD, Correa M: **Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine**. *Behav Brain Res* 2002, **137**:3-25.

73. Nakamura K, Hikosaka O: **Role of dopamine in the primate caudate nucleus in reward modulation of saccades**. *J Neurosci* 2006, **26**:5360-5369.

74. • Mazzoni P, Hristova A, Krakauer JW: **Why don't we move faster? Parkinson's disease, movement vigor, and implicit motivation**. *J Neurosci* 2007, **27**:7105-7116 Patients with Parkinson's disease exhibit bradykinesia, that is, they move more slowly than healthy controls. This paper suggests that this effect is not from the obvious cause of an altered speed-accuracy tradeoff, but rather because the patient's 'motivation to move', defined in a careful way that resonates well with Niv et al's [71] notion of vigor, is reduced..

75. Pessiglione M, Schmidt L, Draganski B, Kalisch R, Lau H, Dolan RJ, Frith CD: **How the brain translates money into force: a neuroimaging study of subliminal motivation**. *Science* 2007, **316**:904-906.

76. Shidara M, Richmond BJ: **Differential encoding of information about progress through multi-trial reward schedules by three groups of ventral striatal neurons**. *Neurosci Res* 2004, **49**:307-314.

77. Glimcher PW: *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics (Bradford Books)*. The MIT Press; 2004.

78. Montague PR: **Neuroeconomics: a view from neuroscience**. *Funct Neurol* 2007, **22**:219-234.

79. Daw ND, Kakade S, Dayan P: **Opponent interactions between serotonin and dopamine**. *Neural Netw* 2002, **15**:603-616.

80. Ungless MA, Magill PJ, Bolam JP: **Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli**. *Science* 2004, **303**:2040-2042.

81. Coizet V, Dommett EJ, Redgrave P, Overton PG: **Nociceptive responses of midbrain dopaminergic neurones are modulated**

by the superior colliculus in the rat. *Neuroscience* 2006, **139**:1479-1493.

82. Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS: **Temporal difference models describe higher-order learning in humans**. *Nature* 2004, **429**:664-667.

83. Jensen J, Smith AJ, Willeit M, Crawley AP, Mikulis DJ, Vitcu I, Kapur S: **Separate brain regions code for salience vs. valence during reward prediction in humans**. *Hum Brain Mapp* 2007, **28**:294-302.

84. Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA: **Tracking the hemodynamic responses to reward and punishment in the striatum**. *J Neurophysiol* 2000, **84**:3072-3077.

85. Menon M, Jensen J, Vitcu I, Graff-Guerrero A, Crawley A, Smith MA, Kapur S: **Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: effects of dopaminergic modulation**. *Biol Psychiatry* 2007, **62**:765-772.

86. Frank MJ, Seeberger LC, O'Reilly RC: **By carrot or by stick: cognitive reinforcement learning in Parkinsonism**. *Science* 2004, **306**:1940-1943.

87. Seymour B, Daw N, Dayan P, Singer T, Dolan R: **Differential encoding of losses and gains in the human striatum**. *J Neurosci* 2007, **27**:4826-4831.

88. Reynolds SM, Berridge KC: **Fear and feeding in the nucleus accumbens shell: rostrocaudal segregation of GABA-elicited defensive behavior versus eating behaviour**. *J Neurosci* 2001, **21**:3261-3270.

89. Reynolds SM, Berridge KC: **Positive and negative motivation in nucleus accumbens shell: bivalent rostrocaudal gradients for GABA-elicited eating, taste "liking"/"disliking" reactions, place preference/avoidance, and fear**. *J Neurosci* 2002, **22**:7308-7320.

90. O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C: **Abstract reward and punishment representations in the human orbitofrontal cortex**. *Nat Neurosci* 2001, **4**:95-102.

91. McNaughton N, Corr PJ: **A two-dimensional neuropsychology of defense: Fear/anxiety and defensive distance**. *Neurosci Biobehav Rev* 2004, **28**:285-305.

92. Moutoussis M, Williams J, Dayan P, Bentall RP: **Persecutory delusions and the conditioned avoidance paradigm: towards an integration of the psychology and biology of paranoia**. *Cognit Neuropsychiatry* 2007, **12**:495-510.

93. Kim H, Shimojo S, O'Doherty JP: **Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain**. *PLoS Biol* 2006, **4**:e233.

94. Holland P, Gallagher M: **Amygdala circuitry in attentional and representational processes**. *Trends Cogn Sci* 1999, **3**:65-73.

95. Cohen JD, McClure SM, Yu AJ: **Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration**. *Philos Trans R Soc Lond B Biol Sci* 2007, **362**:933-942.

96. • Wittmann BC, Bunzeck N, Dolan RJ, Düzel E: **Anticipation of novelty recruits reward system and hippocampus while promoting recollection**. *Neuroimage* 2007, **38**:194-202 Novelty plays a crucial role in learning, and has also been associated with the dopamine system via the construct of bonuses [98]. This paper is part of a series of studies into the coupling between novelty, putatively dopaminergic midbrain activation and hippocampal-dependent long-term memory. Along with their intrinsic interest, the results of these experiments suggest that this form of learning can provide evidence about the characteristics of signals associated with RL..

97. Bunzeck N, Düzel E: **Absolute coding of stimulus novelty in the human substantia nigra/VTA**. *Neuron* 2006, **51**:369-379.

98. Kakade S, Dayan P: **Dopamine: generalization and bonuses**. *Neural Netw* 2002, **15**:549-559.

99. Yoshida W, Ishii S: **Resolution of uncertainty in prefrontal cortex**. *Neuron* 2006, **50**:781-789.

100. Matsumoto M, Matsumoto K, Tanaka K: **Effects of novelty on activity of lateral and medial prefrontal neurons**. *Neurosci Res* 2007, **57**:268-276.

101. Herry C, Bach DR, Esposito F, Salle FD, Perrig WJ, Scheffler K, Lthi A, Seifritz E: **Processing of temporal unpredictability in human and animal amygdale**. *J Neurosci* 2007, **27**:5958-5966.

102. Li J, McClure SM, King-Casas B, Montague PR: **Policy adjustment in a dynamic economic game**. *PLoS ONE* 2006, **1**:e103.

103. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ: **Cortical substrates for exploratory decisions in humans**. *Nature* 2006, **441**:876-879.

104. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS: **Learning the value of information in an uncertain world**. *Nat Neurosci* 2007, **10**:1214-1221.

105. Lee HJ, Youn JM, O MJ, Gallagher M, Holland PC. **Role of substantia nigra-amygdala connections in surprise-induced enhancement of attention**. *J Neurosci* 2006, **26**:6077–6081.

106. Holland PC, Gallagher M: **Different roles for amygdala central nucleus and substantia innominata in the surprise-induced enhancement of learning**. *J Neurosci* 2006, **26**:3791-3797.

107. Camerer C: **Behavioral Game Theory: Experiments in Strategic Interaction** Princeton, NJ: Princeton University Press; 2003.

108. Kuhnen CM, Knutson B: **The neural basis of financial risk taking**. *Neuron* 2005, **47**:763-770.

109. Dreher JC, Kohn P, Berman KF: **Neural coding of distinct statistical properties of reward information in humans**. *Cereb Cortex* 2006, **16**:561-573.

110. Tanaka SC, Samejima K, Okada G, Ueda K, Okamoto Y, Yamawaki S, Doya K: **Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics**. *Neural Netw* 2006, **19**:1233-1241.

111. Tobler PN, O'Doherty JP, Dolan RJ, Schultz W: **Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems**. *J Neurophysiol* 2007, **97**:1621-1632.

112. Preuschoff K, Bossaerts P, Quartz SR: **Neural differentiation of expected reward and risk in human subcortical structures**. *Neuron* 2006, **51**:381-390.

113. Tobler PN, Fletcher PC, Bullmore ET, Schultz W: **Learning-related human brain activations reflecting individual finances**. *Neuron* 2007, **54**:167-175.

114. Niv Y, Joel D, Meilijson I, Ruppin E: **Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors**. *Adaptive Behavior* 2002, **10**:5-24.

115. Mihatsch O, Neuneier R: **Risk-sensitive reinforcement learning**. *Mach Learn* 2002, **49**:267-290.

116. Byrne R: **Mental models and counterfactual thoughts about what might have been**. *Trends Cogn Sci* 2002, **6**:426-431.

117. Lohrenz T, McCabe K, Camerer CF, Montague PR: **Neural signature of fictive learning signals in a sequential investment task**. *Proc Natl Acad Sci U S A* 2007, **104**:9493-9498.

118. Bell D: **Regret in decision making under uncertainty**. *Oper Res* 1982, **30**:961-981.

119. Loomes G, Sugden R: **Regret theory: an alternative theory of rational choice under uncertainty**. *Econ J* 1982, **92**:805-824.

120. Breiter HC, Aharon I, Kahneman D, Dale A, Shizgal P: **Functional imaging of neural responses to expectancy and experience of monetary gains and losses**. *Neuron* 2001, **30**:619-639.

121. Coricelli G, Critchley HD, Joffily M, O'Doherty JP, Sirigu A, Dolan RJ: **Regret and its avoidance: a neuroimaging study of choice behaviour**. *Nat Neurosci* 2005, **8**:1255-1262.

122. De Martino B, Kumaran D, Seymour B, Dolan RJ: **Frames, biases, and rational decision-making in the human brain**. *Science* 2006, **313**:684-687.

123. Fehr E, Camerer CF: **Social neuroeconomics: the neural circuitry of social preferences**. *Trends Cogn Sci* 2007, **11**:419-427.

124. Chiu PH, Kayali MA, Kishida KT, Tomlin D, Klinger LG, Klinger MR, Montague PR: **Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism**. *Neuron* 2008, **57**:463-473.

125. Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E: **Diminishing reciprocal fairness by disrupting the right prefrontal cortex**. *Science* 2006, **314**:829-832.

126. Shoham Y, Powers R, Grenager T: **If multi-agent learning is the answer, what is the question?** *Artif Intell* 2007, **171**:365-377.

127. Gmytrasiewicz P, Doshi P: **A framework for sequential planning in multi-agent settings**. *J Artif Intell Res* 2005, **24**:49-79.

128. Breland K, Breland M: **The misbehavior of organisms**. *Am Psychol* 1961, **16**:681-684.

129. Talmi D, Seymour B, Dayan P, Dolan RJ: **Human Pavlovian-instrumental transfer**. *J Neurosci* 2008, **28**:360-368.

130. El-Amamy H, Holland PC: **Dissociable effects of disconnecting amygdala central nucleus from the ventral tegmental area or substantia nigra on learned orienting and incentive motivation**. *Eur J Neurosci* 2007, **25**:1557-1567.

131. Murschall A, Hauber W: **Inactivation of the ventral tegmental area abolished the general excitatory influence of Pavlovian cues on instrumental performance**. *Learn Mem* 2006, **13**:123-126.

132. Graeff FG, Guimares FS, Andrade TGD, Deakin JF: **Role of 5-HT in stress, anxiety, and depression**. *Pharmacol Biochem Behav* 1996, **54**:129-141.

133. Soubrié P: **Reconciling the role of central serotonin neurons in human and animal behaviour**. *Behav Brain Sci* 1986, **9**:364.

134. Dayan P, Huys QJM: **Serotonin, inhibition, and negative mood**. *PLoS Comput Biol* 2008, **4**:e4.

135. Blanchard DC, Blanchard RJ: **Ethoexperimental approaches to the biology of emotion**. *Annu Rev Psychol* 1988, **39**:43-68.

136. Mobbs D, Petrovic P, Marchant JL, Hassabis D, Weiskopf N, Seymour B, Dolan RJ, Frith CD: **When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans**. *Science* 2007, **317**:1079-1083.

137. Dayan P, Niv Y, Seymour B, Daw ND: **The misbehavior of value and the discipline of the will**. *Neural Netw* 2006, **19**:1153-1160.

138. • Frank MJ, Samanta J, Moustafa AA, Sherman SJ: **Hold your horses: impulsivity, deep brain stimulation, and medication in Parkinsonism**. *Science* 2007, **318**:1309-1312 An earlier paper by the same group [141] suggested that the subthalamic nucleus (STN) might play a crucial role in slowing down behavior when a choice has to be made between strongly appetitive options, so the absolute best one can be chosen without impulsive early choices. This very clever study uses two different treatments for Parkinson's disease (dopaminergic therapy and deep brain stimulation of the STN) together with a very important prediction/action learning task [86] to test this hypothesis, and the distinction between STN and dopaminergic effects in the striatum..

139. Aron AR, Poldrack RA: **Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus**. *J Neurosci* 2006, **26**:2424-2433.

140. Aron AR, Behrens TE, Smith S, Frank MJ, Poldrack RA: **Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI**. *J Neurosci* 2007, **27**:3743-3752.

141. Frank MJ: **Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making**. *Neural Netw* 2006, **19**:1120-1136.

142. Atallah HE, Lopez-Paniagua D, Rudy JW, O'Reilly RC: **Separate neural substrates for skill learning and performance in the ventral and dorsal striatum**. *Nat Neurosci* 2007, **10**:126-131.

143. Belin D, Everitt BJ: **Cocaine seeking habits depend upon dopamine-dependent serial connectivity linking the ventral with the dorsal striatum**. *Neuron* 2008, **57**:432-441.

144. Joel D, Weiner I: **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* 2000, **96**:451-474.

145. Haber SN, Fudge JL, McFarland NR: **Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum**. *J Neurosci* 2000, **20**:2369-2382.

146. Barto A, Mahadevan S: **Recent advances in hierarchical reinforcement learning**. *Discrete Event Dynamic Systems* 2003, **13**:341-379.

147. Botvinick MM: **Hierarchical models of behavior and prefrontal function**. *Trends Cogn Sci* 2008, **12**:201-208.

148. Bouton ME: **Context, ambiguity, and unlearning: sources of relapse after behavioral extinction**. *Biol Psychiatry* 2002, **60**:322-328.

149. Redish AD, Jensen S, Johnson A, Kurth-Nelson Z: **Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling**. *Psychol Rev* 2007, **114**:784-805.

150. Montague PR, McClure SM, Baldwin PR, Phillips PEM, Budygin EA, Stuber GD, Kilpatrick MR, Wightman RM: **Dynamic gain control of dopamine delivery in freely moving animals**. *J Neurosci* 2004, **24**:1754-1759.

151. Tobler PN, Dickinson A, Schultz W: **Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm**. *J Neurosci* 2003, **23**:10402-10410.

152. Tobler PN, Fiorillo CD, Schultz W: **Adaptive coding of reward value by dopamine neurons**. *Science* 2005, **307**:1642-1645.

153. Preuschoff K, Bossaerts P: **Adding prediction risk to the theory of reward learning**. *Ann NY Acad Sci* 2007, **1104**:135-146.

154. Lammel S, Hetzel A, Hckel O, Jones I, Liss B, Roeper J: **Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system**. *Neuron* 2008, **57**:760-773.

155. O'Reilly RC, Frank MJ: **Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia**. *Neural Comput* 2006, **18**:283-328.

156. Schultz W, Dayan P, Montague PR: **A neural substrate of prediction and reward**. *Science* 1997, **275**:1593-1599.

157. Gibbon J, Malapani C, Dale C, Gallistel C: **Toward a neurobiology of temporal cognition: advances and challenges**. *Curr Opin Neurobiol* 1997, **7**:170-184.

158. Day JJ, Roitman MF, Wightman RM, Carelli RM: **Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens**. *Nat Neurosci* 2007, **10**:1020-1028.

159. Hyland BI, Reynolds JNJ, Hay J, Perk CG, Miller R: **Firing modes of midbrain dopamine cells in the freely moving rat**. *Neuroscience* 2002, **114**:475-492.

160. Pan WX, Schmidt R, Wickens JR, Hyland BI: **Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network**. *J Neurosci* 2005, **25**:6235-6242.