



Review

Cite this article: Daw ND, Dayan P. 2014

The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. B* **369**: 20130478.

<http://dx.doi.org/10.1098/rstb.2013.0478>

One contribution of 18 to a Theme Issue 'The principles of goal-directed decision-making: from neural mechanisms to computation and robotics'.

Subject Areas:

theoretical biology, cognition

Keywords:

reinforcement learning, model-based reasoning, model-free reasoning, striatum, orbitofrontal cortex, Monte Carlo tree search

Author for correspondence:

Peter Dayan

e-mail: dayan@gatsby.ucl.ac.uk

The algorithmic anatomy of model-based evaluation

Nathaniel D. Daw¹ and Peter Dayan²

¹Department of Psychology and Center for Neural Science, New York University, 4 Washington Place Suite 888, New York, NY 10003, USA

²Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London WC1N 3AR, UK

NDD, 0000-0001-5029-1430; PD, 0000-0003-3476-1839

Despite many debates in the first half of the twentieth century, it is now largely a truism that humans and other animals build models of their environments and use them for prediction and control. However, model-based (MB) reasoning presents severe computational challenges. Alternative, computationally simpler, model-free (MF) schemes have been suggested in the reinforcement learning literature, and have afforded influential accounts of behavioural and neural data. Here, we study the realization of MB calculations, and the ways that this might be woven together with MF values and evaluation methods. There are as yet mostly only hints in the literature as to the resulting tapestry, so we offer more preview than review.

1. Introduction

Animals and humans often occupy environments in which there is a series of distinct states which are linked by possibly stochastic temporal dynamics or transitions. In turn, these states are associated, again possibly stochastically, with outcomes that can be appetitive or aversive. In order to choose actions to obtain the appetitive outcomes and avoid the aversive ones, it is typically necessary to make predictions about the net utility of the long-run delivery of these outcomes, given a particular candidate action.

There are widely believed to be at least two structurally different methods for making such predictions. In the context of reinforcement learning, they are referred to as model-free (MF) and model-based (MB) reasoning [1,2], with the former being retrospective and the latter being prospective. Other fields that make somewhat similar distinctions use different labels, such as habitual and goal-directed [3–5]; not to mention yet more distant dichotomies such as type I and type II reasoning [6,7].

The MB prediction of long-run future utility starting from a state x is based explicitly on (i) a sum over the whole set (typically in the form of an expanding tree) of future states that the subject's (typically learned) model tells it to expect to encounter following that state and (ii) the outcomes that the model tells it to expect to receive at those potential future states, each endowed with a utility that is assessed by a second aspect of the model. By contrast, the MF prediction depends only implicitly on these quantities. It is underpinned by the observation that since long-run utility is at stake, the prediction made at state x should be consistent with the equivalent predictions made at the neighbouring states which the subject visits following x [2,8]. On average, the only difference should be the utility provided at x itself. Thus, the MF prediction arises from learning from past encounters of state x to minimize untoward inconsistencies. Even more extreme MF systems directly couple observations to decisions, for instance using the inconsistencies to learn appropriate actions or action sequences [9–11].

The differences between MB and MF learning matter. First, they are known to have quite disparate computational and statistical properties [1,4]. MF predictions are computationally simple because they typically depend on little more than a feed-forward mapping of state to predicted future utility. However, they are statistically infelicitous, because adjusting long-run predictions by reducing local inconsistencies between neighbouring states' predictions fails

to take account at every juncture of all the information that is known about the values of more distant states. Instead, this approach, known as bootstrapping, uses the value currently predicted at x 's successor state as a stand-in for the value of all rewards to follow. This biases the estimate of x , particularly at the onset of learning when all the values are poorly known. Conversely, MB methods are computationally ruinous, because there are usually very many possible future states in a long series of transitions. However, these methods are statistically straightforward, because learning can be temporally completely local.

Second, MB and MF methods have quite distinct psychological properties, which underpin a venerable debate about their relative importance [4,5,12–16]. Consider what happens when the environment changes, so that either the utility associated with a particular outcome or some state transition is different from its value during learning. The former could happen if the animal learned when hungry, but is now thirsty; the latter if a path in a maze is newly blocked or opened. Such adjustments can affect the utility that should be predicted at many states antecedent to where the change occurred. Because MF methods are based on the local reduction of prediction errors, there will be no update to predictions made at states far from the change until paths collectively leading there have actually been experienced. By contrast, MB predictions at distant states can alter straightaway given observation of just the changes, because these predictions are constructed on the fly using current information. In normal human subjects, it seems that choice can depend on both MB and MF methods, with the balance being tipped to the latter by a number of manipulations such as cognitively demanding dual tasks [17] or stress [18].

These rather clear psychological distinctions have, at least to a degree, helped elucidate the neural circuitry involved in MB and MF reasoning. Unit recordings and other studies have suggested that appetitive MF learning is supported by prediction errors carried by dopaminergic neurons in the midbrain, driving plasticity at their targets in striatum and elsewhere [19–21]. Meanwhile, lesion-based studies in rodents suggest that there is also a particular role for the dorsolateral striatum in MF control and possibly the central nucleus of the amygdala in prediction [22–27]. Equally, ventral prefrontal areas, the dorsomedial striatum, and possibly the basolateral nucleus of the amygdala are implicated in MB prediction and control [22,23,27]. These localizations have also been partly verified in a set of neuroimaging experiments in humans [28–31].

However, although these studies, and a wealth of others, are revealing as to the regions of the brain that are involved, much less is known about the actual mechanisms supporting MB control in the brain. To a great extent, this is because these experiments predominantly turn on simply detecting MB evaluation (e.g. testing whether choices or neural signals adjust appropriately when the environment changes) but tend, as a group, not to address the within-trial processes by which such computations occur. Such an experimental approach is well matched to theories of MB planning in psychology, which tend to be at Marr's computational level. A more detailed, algorithmic or process-level account might serve as a framework for interpreting the neural substrate. In reviewing different computational approaches to MB evaluation, this review aims to lay out a set of possible directions for filling in these gaps.

In so doing, and in contrast to existing psychological models (and to some extent, experimental tasks) that consider MB computation in small and tractable state spaces, we focus on the problems that arise in more realistic domains when the number of states or trajectories is large [4,32–34]. Then, it is not usually possible to perform a full MB calculation, i.e. to enumerate the full tree of future states and evaluate simulated net utilities. This motivates a search for alternatives. One prominent example is the substitution of MF for MB estimates at parts of the tree, thus obviating the requirement for full MB search. Doing this judiciously is a substantial meta-control problem [35–38] that we do not yet know how to solve. However, such interactions, if understood more systematically, might provide us with a basis for understanding various puzzling interactions between MB and MF systems that appear to be suggested by recent neuroscientific and behavioural data [39–43].

Here, we consider MB prediction from various algorithmic angles, discussing key methods and problems. We do not intend to be computationally, psychologically or neurally comprehensive, but instead mostly point to some possibilities that merit further examination. In §2, we define the problem for MB prediction in a Markov domain. In §3, we consider shortcuts of various sorts and their potential origin in learning. Finally, in §5, we touch on some of the neural implications and relationships for these notions.

2. Model-based prediction

To begin, we lay out the problem of predicting long-run utility using a model. To streamline the discussion, we adopt two formal simplifications. First, much of the recent literature on MB methods has concerned control—i.e. the choice of trajectories of actions in order to optimize long-run utility [2]. However, many of the critical issues are raised by the simpler problem of *prediction* of the long-run utility, and so we initially focus on this. Second, following much work in reinforcement learning (RL), we assume that the problem is expressed in terms of states x that satisfy the Markov property that future states and rewards are conditionally independent of past ones, given the current state. In problems where this fails to hold for the most obvious definition of state, a generic manoeuvre is to define a (larger) space of auxiliary states that *are* Markovian. For instance, a hidden Markov model or a partially observable Markov decision process involves a latent or hidden state whose rendition in the actual observations is partial or confounded. However, although the observations therefore fail to satisfy the Markov property, the belief state, which is the posterior distribution over the latent process state given the observations, does [44]. In this way, the basic framework described below also applies to these settings.

Thus, consider an uncontrolled Markov chain over states $x \in \mathcal{X}$ with transition structure $T_{xy} = P[x(t) = y | x(t-1) = x]$ and expected reward vector $r_x = \mathcal{E}[R(t) | x(t) = x]$. For convenience, we will consider deterministic rewards; the methods can be readily extended to the stochastic case. We will assume that the rewards are bounded.

We take the goal as being to estimate the discounted, long-run future expected reward starting from each state x

$$v_x = \mathcal{E} \left[\sum_{t=0}^{\infty} \gamma^t R(t) | x(0) = x \right], \quad (2.1)$$

where $0 \leq \gamma < 1$ is the discount factor that downweights distant compared with proximal rewards, and the expectation is taken over the stochasticity of the transitions. We can explicitly expand the expectation as

$$v_x = \sum_{t=0}^{\infty} \gamma^t \left(\sum_y P[x(t) = y | x(0) = x] r_y \right), \quad (2.2)$$

and then, noting that $P[x(t) = y | x(0) = x] = \mathcal{T}_{xy}^t$ can be expressed using the t th power of the transition matrix, we can derive two simpler forms

$$v_x = \sum_y [\mathcal{I}_{xy} + \gamma \mathcal{T}_{xy} + \gamma^2 \mathcal{T}_{xy}^2 + \dots] r_y \quad (2.3)$$

and

$$v_x = r_x + \gamma \sum_y \mathcal{T}_{xy} v_y, \quad (2.4)$$

where expression (2.4) is a form of consistency condition between the values of successive states. By constructing vectors $\mathbf{v} = \{v_x\}$ and $\mathbf{r} = \{r_x\}$, we can write equation (2.3) as

$$\mathbf{v} = [\mathcal{I} + \gamma \mathcal{T} + \gamma^2 \mathcal{T}^2 + \dots] \mathbf{r} \quad (2.5)$$

and noting the matrix identity that $[\mathcal{I} + \gamma \mathcal{T} + \gamma^2 \mathcal{T}^2 + \dots] = (\mathcal{I} - \gamma \mathcal{T})^{-1}$ we can write this as

$$\mathbf{v} = (\mathcal{I} - \gamma \mathcal{T})^{-1} \mathbf{r}. \quad (2.6)$$

The vector form of equation (2.4) is

$$\mathbf{v} = \mathbf{r} + \gamma \mathcal{T} \mathbf{v}. \quad (2.7)$$

We sometimes write the true value of \mathbf{v} as \mathbf{v}_* . Assume now that we have estimates $\hat{\mathcal{T}}$ and $\hat{\mathbf{r}}$ of \mathcal{T} and \mathbf{r} . What should the estimate be of \mathbf{v}_* , and how can we actually perform the calculations concerned? Although there are other options, one obvious estimate is $\hat{\mathbf{v}}_* = (\mathcal{I} - \gamma \hat{\mathcal{T}})^{-1} \hat{\mathbf{r}}$. This is called the certainty-equivalent value, because it comes from substituting the estimated values into equation (2.6) as if they were true. In terms of calculations, the predictions then also satisfy two simple relationships:

$$\hat{\mathbf{v}} = \hat{\mathbf{r}} + \gamma \hat{\mathcal{T}} \hat{\mathbf{r}} + \gamma^2 \hat{\mathcal{T}}^2 \hat{\mathbf{r}} + \dots \quad (2.8)$$

which is the approximated vector form of equation (2.5)

$$\hat{\mathbf{v}} = \hat{\mathbf{r}} + \gamma \hat{\mathcal{T}} \hat{\mathbf{v}} \quad (2.9)$$

which is the approximated vector form of equation (2.9). These relationships underpin MB and MF algorithms.

(a) Enumeration and roll-outs

In order to compute values $\hat{\mathbf{v}}$ from an estimated model $\hat{\mathcal{T}}$ and $\hat{\mathbf{r}}$, equation (2.8) invites us to perform either a deterministic calculation or stochastic ‘roll-outs’. The deterministic calculation would realize the steps of the addition in the equation, up to a finite horizon s by which point γ^s is sufficiently small relative to the maximal reward. However, this sort of evaluation mechanism is only credible in the case that there are very few states.

In the more general case of many, or even an infinite number of, states, it is possible to use a stochastic sampling method to focus an approximation \hat{v}_x around any state x [45]. One option is the simple Monte Carlo scheme of sampling trajectories from $\hat{\mathcal{T}}$, along with sample rewards $\hat{\mathbf{r}}$, and then averaging the discounted sum of encountered utilities across the samples. One distinct advantage of this is that it suffices

to have a ‘black box’ simulator—thus even if there is an infinite number of states, so that writing down $\hat{\mathcal{T}}_{xy}$ is impossible, it could still be viable to generate sample transitions.

There are two sources of error in such a scheme. The first is the obvious one of only using a finite number of samples; this implies that the average will be corrupted by noise whose standard deviation will decrease with the inverse square root of the number of samples. The second source of error comes from the finite truncation of the infinite horizon.

Because this sort of stochastic evaluation is focused on a particular starting state x , it is not guaranteed to provide the information necessary to evaluate any other state. The latter is not an accident—it is actually a fundamental aspect of this method (called sparse-sampling; [45]). It is, perhaps surprisingly, possible to evaluate expectations well without enumerating anything like all the states.

(b) Consistency

The alternative to enumeration, calculation and roll-outs is to consider equation (2.9) as suggesting a consistency condition that applies between values of \hat{v}_x and \hat{v}_y when it is possible to make a transition from x to y in one step. Inconsistency, i.e. when the two sides of the equation are not equal, can be a signature for learning.

One standard technique is Bellman evaluation, which consists of starting from any $\hat{\mathbf{v}}(0)$, and then performing the iteration which comes directly from the consistency condition of equation (2.9)

$$\hat{\mathbf{v}}(s+1) = \hat{\mathbf{r}} + \gamma \hat{\mathcal{T}} \hat{\mathbf{v}}(s). \quad (2.10)$$

Because $\hat{\mathcal{T}}$ is a stochastic matrix, it is possible to show that this is a contraction (in an \mathcal{L}_∞ norm), in that

$$\max |\hat{\mathbf{v}}(s+1) - \hat{\mathbf{v}}_*| \leq \gamma \max |\hat{\mathbf{v}}(s) - \hat{\mathbf{v}}_*| \quad (2.11)$$

(taking the maximum over the elements of the vectors) which implies that $\hat{\mathbf{v}}(s)$ converges exponentially quickly to the correct answer [46].

Note that for $\hat{\mathbf{v}}(0) = 0$, this method closely corresponds to the enumeration scheme from equation (2.8), with each iteration corresponding to one step of the addition. Apart from skirting explicit computation of powers of the transition matrix, the recursive evaluation here suggests a different possibility for sample-based evaluation (based on single transitions rather than on full trajectories), described next. In addition, the possibility of initializing the iteration with a set of values $\hat{\mathbf{v}}(0)$ other than zero suggests many possibilities for using MB evaluation to refine pre-existing estimates, discussed below in §3.

Replacing $\hat{v}_x(s)$ by $\hat{v}_x(s+1) = \hat{r}_x + \sum_y \hat{\mathcal{T}}_{xy} \hat{v}_y(s)$ is usually called performing a *backup*, in that information from the *future* of state x is used to change the estimated value of state x itself. Of course, until $\hat{v}_y(s)$ is itself correct, the information that is being backed up is not completely faithful. It is in this sense that enforcing consistency is known as a bootstrapping technique.

One could also imagine simply visiting state x° , then generating a sample state y° from $\hat{\mathcal{T}}_{x^\circ y}$ and using $\hat{r}_{x^\circ} + \hat{v}_{y^\circ}(s)$ as a sample backup. In this case, it is necessary to average over multiple samples to overcome the inherent stochasticity, and so to perform a partial backup

$$\begin{aligned} \hat{v}_{x^\circ}(s+1) &= \hat{v}_{x^\circ}(s) + \epsilon(s)(\hat{r}_{x^\circ} + \gamma \hat{v}_{y^\circ}(s) - \hat{v}_{x^\circ}(s)) \\ \hat{v}_{x \neq x^\circ}(s+1) &= \hat{v}_{x \neq x^\circ}(s). \end{aligned} \quad (2.12)$$

Here, $\epsilon(s)$ is a learning rate; the term it multiplies is called the temporal difference (TD) prediction error [47]. Provided each state x° is visited infinitely often, and some other technical conditions are satisfied, this scheme will converge appropriately: $\lim_{s \rightarrow \infty} \hat{v}(s) = \hat{v}_*$. We discuss the MF use of this rule in §3a.

(c) Roll-outs and consistency

It is possible to combine rollouts and consistency, as in the original sparse-sampling algorithm [45] and varieties of Monte Carlo tree search techniques [48]. In a variant that was really intended for control rather than just for prediction, one idea is to build a tree-structured representation of the value of a subset of states progressively, one node (representing one state's value) at a time. Evaluation begins at the root of the tree, which is the current state. Sample transitions move the state down the tree (with sample rewards also being noted). Any time a leaf of the current tree is reached, a (curtailed) rollout is performed without building any more tree structure; the rewards and values found are backed up in the existing tree along the sample path (using a version of equation (2.12)). Then, a new node is added to the tree at the leaf that has just been evaluated, and a new evaluation commences from the root. Under suitable assumptions about the Markov decision process (MDP), it is possible to bound the expected error following a certain number of roll-outs.

There are two main advantages of making the tree explicit and using back-ups, and one main disadvantage. One advantage is that if the actual transition in the world follows one of the paths in the tree, as is likely given sufficient tree-based evaluation, then less work needs to be done in the next iteration, because part of the tree will remain relevant. A second stems from reducing the variance, because the values at intermediate nodes in the tree that are used as part of the back-ups, have themselves benefited from being averages over a set of stochastic samples on previous trials. The disadvantage is that it is necessary to represent the tree, which can be expensive, at very least in terms of memory.

(d) Discussion

The methods discussed in this section use more or less expensive calculations to make predictions prospectively based on \hat{r} and \hat{T} . Such predictions (and similarly decisions, in the case of control) conform to the key signature of goal-directed behaviour [5] that they immediately reflect changes in either quantity. In particular, consider a new reward vector \hat{r}' (e.g. induced by a change from hunger to satiety) and/or a new transition matrix \hat{T}' (representing different contingencies, such as a rearranged maze). If a subject were to learn about such changes, then MB computations based on the new model would (correctly) change the predicted values straightaway to the new implied values \hat{v}' . This change is missed by conventional MF methods.

It is important to be able to quantify the uncertainty about values such as \hat{v} —in this case depending on a given amount of enumeration, roll-outs or samples. Although doing so precisely is extremely demanding in terms both of prior knowledge about the domain and computation, it is often possible to bound the uncertainty, given relatively coarse extra information such as the range of possible rewards and probabilities. In principle, one would like to exploit this information to target future computation more precisely, as a form

of what is known as meta-control [35,37,38]. We discuss this further in §4.

3. Ameliorating model-based methods

The trouble with the MB evaluation algorithms discussed in §2 is that they pose severe challenges to computation (by requiring multiple iterations or roll-outs) and in some cases additionally to memory (through the size of the tree). This motivates an examination of alternatives that use estimates or approximations to simplify or replace the MB calculations. Two main suggestions have attracted substantial attention: (i) making and storing direct estimates of the long-run utilities of states that can substitute for the complexities of calculation, and (ii) employing forms of temporal abstraction.

(a) Direct value estimates

It was an early triumph of artificial intelligence to show that it is possible to acquire estimates \tilde{v} of the *endpoint* \hat{v} (or even of \mathbf{v}_*) of MB calculation directly from experience of action transitions and rewards without building or searching a model (hence the MF moniker). This involves enforcing consistency between estimates made at successive states, an idea from Samuel [8] that led to the TD learning rule of Sutton [47] mentioned in §2b. It can be seen as substituting actual transitions and utilities sampled from the environment for the simulated experience that we discussed as underpinning various forms of MB evaluation.

That is, the accumulated, discounted sequence of rewards received following the visit to some state is a sample of its value, and, like simulated roll-outs, these samples can be averaged over state visits. In addition, exactly analogous to the sample backups of equation (2.12), a prediction error can be computed from any observed state-reward-state progression, and used to update the estimated value of the earlier of the states. This is the TD method [47] for online value estimation. A variant called TD(λ) interpolates between the two extremes of estimating a state's value in terms of the full sequence of rewards that follows it, versus bootstrapping based only on consistency with the immediate reward and the value of the next state. These sampling methods clearly have the same convergence guarantees as the corresponding MB sample methods.

Given that we can produce estimates of values, what can we do with them? Most obviously, MF quantities \tilde{v} can replace MB ones \hat{v} as estimates (or as their policy-maximizing counterparts for control). Perhaps the more interesting question from our perspective here, however, is whether they can be used in conjunction with the MB schemes in §2, to improve on the performance of either alone [49]. The main idea is that they can provide a starting point for iterative improvement via further MB computation. Thus, for instance, in the recursive Bellman backup scheme of equation (2.10), values $\hat{v}(0)$ can be initialized to \tilde{v} , potentially improving the bootstrapping backups and ultimately speeding convergence, if these starting points are close to \hat{v}_* . Similarly, the iterated sum from equation (2.8) can proceed progressively through a series of terms, but then instead of truncating the sum, terminate with the approximate values $\gamma^s \hat{T}^s \tilde{v}$. Generally, in tree traversal, estimated values can be substituted so as to treat a branch like a leaf. Finally, in just the same way, MB methods using local backups, sample roll-outs or transitions can be

applied locally to improve the existing estimates at any given state x .

Such a combination of MF estimates with MB refinement is motivated by the fact that estimating $\tilde{\mathbf{v}}$ from MF experience is so closely related to some of the methods we discussed for computing $\hat{\mathbf{v}}$ from a model. That both involve accumulating samples, though from different sources, makes it seem natural more freely to intermix both sorts of samples, experiential and model-generated, in updating the same vector. This is the idea behind the Dyna architecture [50], and since has been refined to take better advantage of search trees [51,52]. These considerations suggest that it may make sense for MF learned values and model-derived computed values to share a single value store, which is both updated online from experience and subjected to refinement with MB methods. This store can be seen as caching the conclusions from experience in the world, expensive MB evaluations, and also ersatz experience generated from a model.

Viewed this way, the question becomes whether, and at what states, it is worth spending the computational and memory costs of MB computation to refine pre-existing value estimates, and, conversely, whether learning $\tilde{\mathbf{v}}$ directly will permit computational savings relative to purely MB computation, and at what cost. We consider aspects of this in §3c.

(b) Abstraction

A second approach to ameliorating the problems of evaluating long-run returns involves temporal abstraction. It can apply to MB and MF evaluation; we describe it first for the latter in the context of feature-based representations of states.

In §3a we considered a separate MF prediction \tilde{v}_x for each state x . However, it is more common to represent the states using a set of features, and then to realize $\tilde{\mathbf{v}}$ as a linear function of the features. In these representational schemes, if the value of the y th feature at state x is X_{xy} , then we write $\tilde{v}_x = \sum_y X_{xy} \tilde{w}_y$, where $\tilde{\mathbf{w}}$ are the estimating weights (that then become the target of TD learning). Collectively, this makes

$$\tilde{\mathbf{v}} = \mathcal{X}\tilde{\mathbf{w}}. \quad (3.1)$$

Consider using as a representation \mathcal{X} , an estimate $\tilde{\mathcal{M}}$ of $(\mathcal{I} - \gamma\mathcal{T})^{-1}$ [53,54]. Comparing equations (2.6) and (3.1), it is apparent that if the weights $\tilde{\mathbf{w}} = \mathbf{r}$ were just the immediate rewards, then MF predictions would be \mathbf{v}_* . To put it another way, in matrix $\tilde{\mathcal{M}}$, feature y represents the discounted future occupancy of state y starting from each initial state. This is therefore called the *successor matrix*. The value of state x is a sum over such future occupancies, each weighted by the immediate reward r_y available at the states involved.

It is straightforward to learn estimates of the immediate rewards. The successor matrix $\tilde{\mathcal{M}}$ itself can be estimated via sampling methods that are exactly analogous to those for estimating $\tilde{\mathbf{v}}$ directly [53]. Recall that each row $\tilde{\mathbf{m}}_x$ of the matrix represents the expected, discounted cumulative occupancy over all other states, following visit to some state x . Conversely, each column $\tilde{\mathbf{m}}_{\cdot y}$ counts future (discounted) visits to some successor state y , as a function of different start states. A column of $\tilde{\mathcal{M}}$ is thus equivalent to a value function for a Markov process in which unit reward $r = 1$ is earned for each visit to state y , and zero otherwise, defined by a Bellman equation analogous to equation (2.10). Thus,

in turn, each column of $\tilde{\mathcal{M}}$ can be learned via either of the sampling methods in §2 (trajectory or transition based, or in general, TD(λ)), where visits to state y are counted in place of rewards. In general, TD learning of $\tilde{\mathcal{M}}$ requires updating an entire row of the matrix following the transition from some state x to y , according to the observed transition (1 for y , 0 elsewhere) plus a vector-valued backup of the discounted occupancies over all other states expected following y , i.e. $\gamma\tilde{\mathbf{m}}_y$.

The successor matrix is not only useful for MF evaluation. Because it summarizes and thus replaces the trees, roll-outs or iterations over state visits that are required to produce the MB computations based on equations (2.8) or (2.9), it can save almost all the computation associated with the implied algorithms. All that would formally be required is an MB estimate of the utility of the immediate reward at each state—a quantity that is in any case part of the model.

The way the successor representation aggregates future state occupancies is an extreme example of temporal abstraction. Other related ideas include multiscale temporal modelling [54] (itself generalized in [55]), which involves learning a world model corresponding to multiple steps of the Markov transition dynamics. This comprises powers of the transition matrix (e.g. \mathcal{T}^2 for two steps) and the associated sums of rewards ($\mathbf{r} + \gamma\mathcal{T}\mathbf{r}$). These constitute Markov chains, as do arbitrary mixtures of them; each represents views of the same process at different timescales. In the context of MB evaluation, coarser timescale models can allow for deeper search, in effect by aggregating multiple steps of the world model together. They can be learned in a similar manner to the successor matrix.

Similarly, rather than entire transition matrices, individual sets of state transitions can also be aggregated. This arises mainly in the control case, where the set of one-step actions can be augmented by aggregate actions, known as options, which constitute an extended policy [56,57]. Again, an option includes both transition and reward models that aggregate the consequences of the extended action sequence. Because these pre-compute a chunk of the tree of future states (and potentially actions), they can be used to take large steps during ‘saltatory’ MB evaluation [56]. However, whereas the consequences of following a particular temporally extended option (in terms of end state and rewards received) are also easy to learn by essentially the same methods as discussed so far, the larger problem of option discovery, i.e. finding a useful set of temporally extended actions, is an area of substantial current research [11,56].

(c) Trade-offs

All these MB and MF methods enjoy asymptotic optimality guarantees in terms of computation or experience or both. However, trade-offs arise pre-asymptotically. Importantly, error, relative to the true values \mathbf{v}_* , has two components. One is due to limitations from evidence, learning or representation: even when computed exactly, $\hat{\mathbf{v}}_*$ will generally not coincide with the true \mathbf{v}_* , because the former is based on an estimated model $\hat{\mathcal{T}}$ and $\hat{\mathbf{r}}$, for instance, itself learned from limited experience with sample transitions and rewards. Such error might be viewed as irreducible, at least at a given stage of learning. Note that if the state space is large or continuous, computations expressed entirely in terms of either MF values or the successor matrix may have different, and potentially favourable,

properties, compared with working with models which would have only to be approximate [46,58,59].

However, value estimates may differ from \hat{v}_* , further owing to computational insufficiencies: either from approximate or inadequate computation in the MB case (e.g. truncating sums, averaging few samples), or, in the MF case, from noise inherent in the way that \tilde{v} and \tilde{M} were themselves computed from experience during learning. In particular, as we have seen, typical methods for learning values or the successor matrix themselves centre on bootstrapping these quantities by sample-based updates analogous to equation (2.12). Such bootstrapping during learning can fall short, for a given amount of evidence, from reaching the final values \hat{v}_* that could have been obtained had a model been learned from the same evidence, and then enumerated or rolled out. In this regime, further refinement of the MF values using MB methods can improve the estimates.

A source of inaccuracy in MF estimates such as \tilde{v} and \tilde{M} that has been of particular interest in a neuroscience setting comes from the case of change in the utilities or transitions. Going back at least to Tolman [16], these manipulations are designed to allow the subject to update the world model (either \hat{r} or \hat{T}) while not giving them experience that would allow standard learning rules to update MF values \tilde{v} or, in some experiments, the accumulated transition information inside \tilde{M} . Under such an experiment, the MF values are biased towards estimates of the old values v instead of the new values v' . Additional computation with the (correct) model could correct this bias.

In particular, insofar as these techniques blend MB evaluation with pre-computed steps, they might or might not pass as MB in the standard laboratory tests. For instance, predictions based on the successor matrix could adjust to new reward contingencies immediately, if the estimates of r in $\tilde{v} = \tilde{M}r$ are replaced by r' . This would not be the case for weights learned to an arbitrary feature matrix \mathcal{X} . Furthermore, it would apply only to manipulations of reward values; changes in the transition contingencies that would alter \tilde{M} would still lead to errors if the successor matrix were not itself relearned. Thus, behaviour produced by the successor representation is not truly goal-directed, under the strict definition [60] of being sensitive to both the action–outcome contingency (\hat{T}) and the outcome value (\hat{r}): it would pass laboratory tests for the second, but not the first, owing to the way \tilde{M} summarizes the aggregated long-run transitions. Similarly, whether values computed from multi-timescale models or options are able to immediately adjust to reward or transition changes would depend on whether or not these reflect rewards or transitions that are already cached inside their pre-computed aggregates (or, potentially, whether or not the way that the cached quantities might have changed since they were learned can be predicted or is known). The fact that temporally extended options, for instance, contain a model of their aggregate reward consequences, is the basis for a suggestion [10,11] that such action chunking (rather than classical MF value learning) might give rise to the experimental phenomena of habits.

As for whether, or at what states, to target additional computation, one view is that this depends on assessing the uncertainty and/or bias in the existing value estimates, which generally can be done only approximately [61–63]. Locally, this involves tracking uncertainty about the rewards or transitions both by taking account of what has been

observed about them and how they are likely to have changed. The challenge here is that the values at different states are coupled to one another through the transitions, so localized uncertainty or change in the model at any particular state has effects on the value (and uncertainty about the value) at many other states antecedent to it via correlations that it is too expensive to maintain. Heuristics, like prioritized sweeping [64], target MB updates towards states likely to have been affected by recently learned examples.

In theorizing about biological RL [4,32,33], information about uncertainty has been proposed to explain, for example, why in experiments, animals favour MB computation early in training (when bootstrapped values are uncertain and there is likely value to MB computation), but MF values dominate following over-training on a stable task (when there is little uncertainty in any case) [65].

(d) Control

As mentioned in §2, we have so far concentrated on prediction rather than control. Control is typically modelled by a Markov decision process in which the agent chooses an action at each state and the successor state depends jointly on the state and action. The critical extra complexity of control is that the *policy* (which is the mapping from states to possible actions) of the agent has to be optimized during learning. Applying any particular policy reduces a Markov decision process back to a Markov chain, by inducing a policy-dependent state–state transition matrix \mathcal{T} of the sort considered in §2. This implies that the values v (and also the state–action version of the values, known as Q values; [66]) both depend on the policy—more directly, because the value of a state depends in part on the actions taken at all subsequent states. An optimal policy is defined as one that jointly maximizes value at all the states.

The requirement for control leads to the extra problem of exploration versus exploitation [35,48,67]. That is, to the extent that uncertainty can be measured, the value of reducing it can be quantified, again at least approximately, in terms of obtained reward. This is because, in the case of control (i.e. if the values are actually used to make choices), more accurate predictions can earn more reward via producing better choices, though only to the extent the value estimates were unreliable in the first place. In this way, improving uncertain estimates can justify choosing actions that may reveal new information. The same trade-off also governs to what extent it is worth spending computational resources refining existing value estimates via MB evaluation.

The relationship between a Markov decision process and a Markov chain implies that most of the algorithmic methods can be adapted to control. For instance, there is a maximizing variant of the consistency condition in equation (2.9) that leads to a maximizing variant of the MB tree backup [68] and of MF value learning [66], both of which produce the value function associated with the optimal policy. While it is possible to define an analogous optimal successor representation (the successor matrix induced by the optimal policy), this cannot be learned in a simple MF way, separate from the optimal values. Instead, the successor representation and multi-timescale models fit more naturally with a different family of control approaches, which include policy iteration in the MB case and the actor–critic for MF learning [9,59]. These work by adopting some candidate policy,

learning its associated value function and using these values to improve the policy.

The issue of policy dependence also plays an important role in Monte Carlo tree search, because the tree being rolled out depends on the assumed policy, which is also what the tree search is optimizing. There is a Monte Carlo tree search mechanism that tries out different actions according to a measure of their likely worth, but taking the potential benefits of exploration into account [48]. It distinguishes between an (approximately) optimal policy that is being captured by the parts of the tree that have already been built, and a roll-out policy which is a default that is applied at so-far-unexpanded leaves of the tree, and which is almost always suboptimal but provably sufficient for this purpose. For the case of prediction alone the roll-out policy is, in effect, what it is sought to evaluate, and so the prediction tree could have been built more aggressively, adding more than just a single node per step at the root.

With respect to MF methods, we have seen that caching quantities that depend on the rewards and transitions inside a value function \bar{v} , the successor matrix \tilde{M} or other temporally extended representations can cause characteristic errors in prediction when the model changes. The fact that these quantities are both also policy-dependent (and also that the policy itself is cached in methods like the actor-critic) is another similar source of error. Thus, if the policy changes at a state, this should typically induce value and policy changes at other states. However, MF methods may not update these all consistently. Notably, although in the prediction case we suggested that the successor representation behaves like MB reinforcement learning with respect to changes in the rewards r , though not the contingencies \mathcal{T} , in the full control case, the successor representation also fail to adjust properly to some changes in rewards alone. This is because in the control case, changes in rewards r can induce changes in the optimal policy, which in turn affect the effective transition matrix, \mathcal{T} , making it different from the one cached inside \tilde{M} .

4. Neural realizations

We have discussed a number of variants of MB computation, focusing on the case of the prediction of long-run utility. We suggested that in problems of realistic size, MB methods are unlikely to be able to function in isolation, but might instead fall back on MF estimates, at least to some extent. Prediction learning is central to Pavlovian or classical conditioning, for which actions are elicited purely accordance with the predictions rather than because of their contingent effects on the outcomes. However, MB and MF computational issues also arise for the more significant case of instrumental or operant conditioning, in which it is necessary and possible to optimize the choice of actions to maximize rewards and minimize punishments. Indeed, most psychological and neuroscience studies of the two sorts of methods have been conducted in the instrumental case.

Substantial data distinguishing MB and MF instrumental control have come from brain lesion experiments in the face of a reward devaluation probe. Collectively, these appear to demonstrate a reasonably clean double dissociation between two different networks supporting MB and MF control, such that animals reliably behave in one or

other fashion given lesions to specific, different, areas. Instrumental MB and MF behaviour appear to be centred on distinct areas of striatum: dorsomedial and dorsolateral, respectively [22,27,69–71]. Areas of striatum are interconnected with associated regions of cortex via topographic ‘loops’ linking the two structures, via additional basal ganglia nuclei and the thalamus [72–74]. At least in the case of MB behaviour, lesions implicate the entire loop, cortical (prelimbic) and thalamic (mediodorsal) regions affecting MB control [23,75,76]. There is also a key role for the orbitofrontal cortex in the flexible assessment of outcome worth that is essential for MB evaluation [24–26,77]. By contrast, cortical and thalamic elements of a hypothetical parallel MF loop through dorsolateral striatum have yet to be demonstrated. Although the results are less clean cut, neuroimaging in humans has revealed potential counterparts to some of these areas [29–31,69,78,79].

A third subregion of striatum, the ventral part (also known as nucleus accumbens) is more closely associated with Pavlovian rather than instrumental aspects of behaviour (e.g. prediction rather than control). It is well connected with parts of the amygdala, and also the orbitofrontal cortex, which are also implicated in Pavlovian behaviour. Importantly, as we discussed, predictions (as embodied in various Pavlovian behaviours) may themselves, in principle, be computed in either an MB or MF fashion. This distinction has not been nearly as carefully worked out behaviourally or computationally in the Pavlovian case [80]. Nevertheless, there does appear to be some behavioural and neural dissociation between MB- and MF-like Pavlovian behaviours, with the shell region of the accumbens, orbitofrontal cortex and the basolateral nucleus of the amygdala supporting the former (such as specific Pavlovian–instrumental transfer and identity unblocking), and the core of the accumbens and the central nucleus of the amygdala supporting the latter (general Pavlovian instrumental transfer) [22,27,81].

Further, the predictions that feed into Pavlovian actions, whether MB or MF, may also affect instrumental actions, though it is not yet clear exactly how or to what extent. Ventral areas of striatum may be implicated in the performance, if not the acquisition, of instrumental behaviour, consistent with an important role of Pavlovian processes in motivation and regulation of behavioural vigour in Pavlovian to instrumental transfer paradigms [5,82–85]. Perhaps for a similar reason, infralimbic cortex (which is more associated with ventral striatum than with dorsolateral) is also involved in habitual instrumental behaviour [23]. In addition, recent results with disconnection lesions suggest that the basolateral and central nuclei of the amygdala, in communication with dorsomedial and dorsolateral striatum, are required for acquiring MB and MF instrumental behaviours, respectively [86,87].

Collectively, these results are evidence for the mutual independence and disjoint neural realizations of these two sorts of valuation and control. This makes it hard to see how both MB and MF behaviours could be supported by a single, shared value store, as in the most literal implementation of Dyna-like schemes that we discussed in §2a. However, the apparent independence under insult does not rule out the possibility that they interact in the intact brain. Indeed, results from human imaging experiments suggest that the reward prediction errors that are thought to drive online MF learning, are themselves sensitive to MB values [14,39]. This may reflect an MB system training or refining

MF values, as in some of the various methods discussed in §3a. Behavioural results in humans, from tasks involving the combination of instructed and experienced values [40] and those involving retrospective reevaluation [42], may also be consistent with the possibility that the MF system is trained by the MB one.

We have also seen various ways in which MB learning and evaluation can be conducted, via samples, using techniques entirely parallel to the temporal-difference methods most associated with MF learning. There is a fairly well-accepted neural mechanism for performing appetitive MF learning in which the phasic activity of dopamine neurons conveys a TD prediction error to its targets, particularly affecting activity and plasticity at corticostriatal synapses [19–21,88]. It is tempting to suggest that, whether operating on the same synapses (as in Dyna [50]) or on parallel ones in, say, adjacent corticostriatal loops, samples of simulated experience from an MB system could use exactly the same mechanism to update values, leveraging or perhaps replicating, the brain's MF circuitry in the service of MB evaluation. The sample trajectories themselves might be produced by mechanisms such as the replay or pre-play of patterns of hippocampal activity expressing recent or possible paths through an environment [89–92], which is known to be coordinated with cortical and striatal activity [93,94].

We noted in §3c,d that the successor matrix used as a state representation for an MF scheme can adjust immediately to changes in reward contingencies (if predictions are based on $\tilde{M}r'$ rather than on $\tilde{M}r$) though not changes in transitions (which require an alteration to \tilde{M} which can only happen in an MF manner through suitable new real or simulated experience). This can be seen as exploiting TD learning to produce an MF method with some of the features of MB control. Other state representations might also afford useful generalizations, sharing the fruits of learning about one set of states with other sets. Unfortunately, we know rather little about the nature and evolution of cortical representations of states that underpin MF (and MB) predictions, although it is widely believed that these representations do capture and represent explicitly statistical regularities and predictabilities in their input [95–97]. It is indeed a popular notion that the cortex builds a sophisticated representation of the environment that should then afford efficient, hierarchical, MB and/or MF control, and indeed appropriate generalization between related tasks and subtasks. How this works is presently quite unclear.

All of these schemes, from Dyna to the successor representation, that repurpose the MF prediction error to serve MB ends predict dopaminergic involvement in both sorts of valuation. It is reasonably well demonstrated that dopamine does indeed affect MF learning [98–100]; however, there is as yet only mixed evidence for the relationship between dopamine and MB learning [31,43,80,101,102].

Finally, we note the new meta-control problem that has arisen from the need to arbitrate and integrate the multiple methods for making predictions and decisions that we have been discussing [36,103]. One approach to this that has been particularly prominent in the case of the control of working memory (in order to create the sort of history representation that turns a partially observable Markov decision problem into the sort of standard Markovian one we have considered here) is to treat the meta-control decisions as being just like external decisions. This would

be realized by the same underlying circuitry (albeit potentially different regions of structures such as the striatum and the cortex) [104–110]. Of course, the potential regress to meta-meta-control problems and beyond needs to be avoided. Such approaches are tied to considerations about what makes meta-control choices appropriate or optimal; new ideas are emerging about this, notably the relationship to various cognitive limitations [111,112].

5. Discussion

From at least the time of Tolman [16], the advantages of MB reasoning have been clear. It offers subjects statistical efficiency and flexibility, i.e. an ability to deploy their latest knowledge in the service of their current goals. There are limitations in this flexibility. It turns out, for instance, to be very difficult for us (though not necessarily scrub-jays [113]) to make MB predictions about motivational or emotional states that we do not currently inhabit [114]. More significant though, and interpretable as a point of attack from the very outset [115], are its computational demands. The trees of future states (and actions) grow most rapidly in even moderate-sized domains, overwhelming the working memory and calculation capacity of subjects. Hierarchical approaches [116,117] are important, but pose their own challenges of provenance and inference.

We therefore reviewed alternative, MF, methods of reasoning, which operate retrospectively based on experience with utilities. Canonical versions of these are less flexible than MB methods because they rely on explicit experience to change their assessments; their use can thus be distinguished from that of MB systems when behaviour fails to change when aspects of the world change. There are various suggestions for how MF values might come to substitute fully [4,32] or partially [33] for MB values, as in habitization [5]. For instance, the relative certainties of the two systems might be weighed with control favouring the less uncertain [4]. Alternatively, the value of the potential information that could be gained by performing MB calculations could be assessed and weighed against the opportunity and/or cognitive cost of the MB calculation [32,33]. Certainly, there is much current emphasis in human and animal studies on finding ways of elucidating the differential engagement of the systems [12,14,28,31,39,118].

We focused on the methods of MB calculation, and the ways that MF values or MF evaluation methods might be embedded in the MB system and that MB behaviour can arise from purportedly MF systems. We saw very close parallels between various versions of each, for instance between methods that explore the MB tree using stochastic simulations in the model versus methods that learn MF values from sampled experience in the world. Indeed, there are many points on the spectrum between MB and MF that current tasks only dimly illuminate [12].

One possibility invited by these refined interactions between MB and MF systems is to consider MB evaluation at a much finer grain. We can envisage a modest set of operations: such things as creating in working memory a node in a search tree; populating it with an initial MF estimate; sampling one or more steps in the tree using a model; backing value information up in the current tree, possibly improving an MF estimate using the resulting model-influenced (though not necessarily purely

MB) prediction error and finally picking an external action. These fine-grain choices lead to various internal (cognitive) or opportunity costs [32,33,119,120], and are plausibly the solution to a meta-control problem (i.e. concerning the approximately optimal control of control) [36,103,112]. This meta-control problem, which itself could have MB and MF elements to its solution, and presumably depends on learning over multiple tasks, is an important focus for future study.

Finally, many precursors of modern ideas about MB and MF planning were early concerns in the first days of the field of artificial intelligence, and grounded a tradition of experimental work in psychology. Some of these ideas underpin the techniques in reinforcement learning that we have been discussing [8,121,122]. Much computational research considered heuristics for tractably building, searching and pruning decision trees, and the use of value functions to assess intermediate position [123–126], to name but a few. In psychology, such algorithms were adopted as models of human behaviour in chess and other planning problems such as the towers of Hanoi or missionaries and cannibals [126–128]. Error rates, reaction times and self-reports in ‘think-aloud’ planning all appear to suggest the usage of particular heuristics to guide decision tree search.

The direct applicability of these search heuristics to computing expected future values in the RL setting is unclear, however, because the methods we have talked about address what can be seen as a wider class of decision problems than chess or the other planning tasks, involving such additional complications as computing expected returns with respect to stochastic transitions (and potentially rewards) and intermediate rewards along trajectories. Nevertheless, heuristic decision tree pruning has recently arisen as an important feature of human behaviour also in modern RL tasks [129], and other insights from this earlier work are starting to permeate modern notions of hierarchical control [56]. It thus seems a ripe time to revisit these models and results, and attempt to understand how they can be made to relate to the theoretical and experimental phenomena reviewed here.

Acknowledgements. We are very grateful to Yael Niv and our many other collaborators on the studies cited who have helped us frame these issues, including Ray Dolan, Sam Gershman, Arthur Guez, Quentin Huys, John O’Doherty, Giovanni Pezzulo, David Silver and Dylan Simon. We thank two anonymous reviewers for very helpful comments. The authors are listed in alphabetical order.

Funding statement. Funding was from the Gatsby Charitable Foundation (P.D.) and a Scholar Award from the McDonnell Foundation (N.D.D.).

References

- Doya K. 1999 What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* **12**, 961–974. (doi:10.1016/S0893-6080(99)00046-5)
- Sutton RS, Barto AG. 1998 *Reinforcement learning: an introduction (adaptive computation and machine learning)*. Cambridge, MA: MIT Press.
- Adams C, Dickinson A. 1981 Actions and habits: variations in associative representations during instrumental learning. In *Information processing in animals: memory mechanisms* (eds N Spear, R Miller), pp. 143–165. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Daw ND, Niv Y, Dayan P. 2005 Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711. (doi:10.1038/nn1560)
- Dickinson A, Balleine B. 2002 The role of learning in motivation. In *Stevens’ handbook of experimental psychology*, vol. 3 (ed. C Gallistel), pp. 497–533. New York, NY: Wiley.
- Kahneman D. 2010 *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Stanovich K, West R. 2002 Individual differences in reasoning: implications for the rationality debate? In *Heuristics and biases: the psychology of intuitive judgment* (eds T Gilovich, D Griffin, D Kahneman), pp. 421–440. Cambridge, UK: Cambridge University Press.
- Samuel A. 1959 Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **3**, 210–229. (doi:10.1147/rd.33.0210)
- Barto AG, Sutton RS, Anderson CW. 1983 Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybernet.* **13**, 834–846. (doi:10.1109/TSMC.1983.6313077)
- Dezfouli A, Balleine BW. 2012 Habits, action sequences and reinforcement learning. *Eur. J. Neurosci.* **35**, 1036–1051. (doi:10.1111/j.1460-9568.2012.08050.x)
- Dezfouli A, Lingawi NW, Balleine BW. 2014 Habits as action sequences: hierarchical action control and changes in outcome value. *Phil. Trans. R. Soc. B* **369**, 20130482. (doi:10.1098/rstb.2013.0482)
- Dolan RJ, Dayan P. 2013 Goals and habits in the brain. *Neuron* **80**, 312–325. (doi:10.1016/j.neuron.2013.09.007)
- Holland PC. 2004 Relations between Pavlovian–instrumental transfer and reinforcer devaluation. *J. Exp. Psychol. Anim. Behav. Process.* **30**, 104–117. (doi:10.1037/0097-7403.30.2.104)
- Simon DA, Daw ND. 2011 Neural correlates of forward planning in a spatial decision task in humans. *J. Neurosci.* **31**, 5526–5539. (doi:10.1523/JNEUROSCI.4647-10.2011)
- Thorndike E. 1911 *Animal intelligence*. New York, NY: Macmillan.
- Tolman E. 1949 There is more than one kind of learning. *Psychol. Rev.* **56**, 144–155. (doi:10.1037/h0055304)
- Otto AR, Gershman SJ, Markman AB, Daw ND. 2013 The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* **24**, 751–761. (doi:10.1177/0956797612463080)
- Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. 2013 Working-memory capacity protects model-based learning from stress. *Proc. Natl Acad. Sci. USA* **110**, 20 941–20 946. (doi:10.1073/pnas.1312011110)
- Barto A. 1995 Adaptive critics and the basal ganglia. In *Models of information processing in the basal ganglia* (eds J Houk, J Davis, D Beiser), pp. 215–232. Cambridge, MA: MIT Press.
- Montague PR, Dayan P, Sejnowski TJ. 1996 A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947.
- Schultz W, Dayan P, Montague PR. 1997 A neural substrate of prediction and reward. *Science* **275**, 1593–1599. (doi:10.1126/science.275.5306.1593)
- Balleine BW. 2005 Neural bases of food-seeking: affect, arousal and reward in corticostriatal limbic circuits. *Physiol. Behav.* **86**, 717–730. (doi:10.1016/j.physbeh.2005.08.061)
- Killcross S, Coutureau E. 2003 Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex* **13**, 400–408. (doi:10.1093/cercor/13.4.400)
- McDannald MA, Jones JL, Takahashi YK, Schoenbaum G. 2013 Learning theory: a driving force in understanding orbitofrontal function. *Neurobiol. Learn. Mem.* **108C**, 22–27.
- McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. 2011 Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J. Neurosci.* **31**, 2700–2705. (doi:10.1523/JNEUROSCI.5499-10.2011)
- McDannald MA, Takahashi YK, Lopatina N, Pietras BW, Jones JL, Schoenbaum G. 2012 Model-based

- learning and the contribution of the orbitofrontal cortex to the model-free world. *Eur. J. Neurosci.* **35**, 991–996. (doi:10.1111/j.1460-9568.2011.07982.x)
27. Yin HH, Ostlund SB, Balleine BW. 2008 Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur. J. Neurosci.* **28**, 1437–1448. (doi:10.1111/j.1460-9568.2008.06422.x)
 28. Gläscher J, Daw N, Dayan P, O'Doherty JP. 2010 States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595. (doi:10.1016/j.neuron.2010.04.016)
 29. Tricomi E, Balleine BW, O'Doherty JP. 2009 A specific role for posterior dorsolateral striatum in human habit learning. *Eur. J. Neurosci.* **29**, 2225–2232. (doi:10.1111/j.1460-9568.2009.06796.x)
 30. Valentin VV, Dickinson A, O'Doherty JP. 2007 Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci.* **27**, 4019–4026. (doi:10.1523/JNEUROSCI.0564-07.2007)
 31. Wunderlich K, Dayan P, Dolan RJ. 2012 Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci.* **15**, 786–791. (doi:10.1038/nn.3068)
 32. Keramati M, Dezfouli A, Piray P. 2011 Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.* **7**, e1002055. (doi:10.1371/journal.pcbi.1002055)
 33. Pezzulo G, Rigoli F, Chersi F. 2013 The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Front. Psychol.* **4**, 92. (doi:10.3389/fpsyg.2013.00092)
 34. Solway A, Botvinick MM. 2012 Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychol. Rev.* **119**, 120–154. (doi:10.1037/a0026435)
 35. Baum EB. 2004 *What is thought?* Cambridge, MA: MIT Press.
 36. Dayan P. 2012 How to set the switches on this thing. *Curr. Opin. Neurobiol.* **22**, 1068–1074. (doi:10.1016/j.conb.2012.05.011)
 37. Russell SJ, Wefald EH. 1991 *Do the right thing: studies in limited rationality*. Cambridge, MA: MIT Press.
 38. Simon HA. 1982 *Models of bounded rationality*. Cambridge, MA: MIT Press.
 39. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. 2011 Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215. (doi:10.1016/j.neuron.2011.02.027)
 40. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ. 2009 Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain Res.* **1299**, 74–94. (doi:10.1016/j.brainres.2009.07.007)
 41. Doll BB, Simon DA, Daw ND. 2012 The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081. (doi:10.1016/j.conb.2012.08.003)
 42. Gershman SJ, Markman AB, Otto AR. 2012 Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* **24**, 751–761.
 43. Takahashi YK, Roesch MR, Wilson RC, Toreson K, O'Donnell P, Niv Y, Schoenbaum G. 2011 Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat. Neurosci.* **14**, 1590–1597. (doi:10.1038/nn.2957)
 44. Kaelbling LP, Littman ML, Cassandra AR. 1998 Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134. (doi:10.1016/S0004-3702(98)00023-X)
 45. Kearns M, Mansour Y, Ng AY. 2002 A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Mach. Learn.* **49**, 193–208. (doi:10.1023/A:1017932429737)
 46. Bertsekas DP, Tsitsiklis JN. 1996 *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.
 47. Sutton R. 1988 Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44.
 48. Kocsis L, Szepesvári C. 2006 Bandit based Monte Carlo planning. In *Machine learning: ECML 2006* (eds J Fürnkranz, T Scheffer, M Spiliopoulou), pp. 282–293. Berlin, Germany: Springer.
 49. Silver D, Sutton RS, Müller M. 2008 Sample-based learning and search with permanent and transient memories. In *Proc. 25th Int. Conf. on Machine Learning*, pp. 968–975. New York, NY: Association for Computing Machinery.
 50. Sutton R. 1990 Integrated architectures for learning, planning, reacting based on approximating dynamic programming. *Proc. Seventh Int. Conf. on Machine Learning*, pp. 216–224. San Francisco, CA: Morgan Kaufmann.
 51. Baxter J, Tridgell A, Weaver L. 2000 Learning to play chess using temporal differences. *Mach. Learn.* **40**, 243–263. (doi:10.1023/A:1007634325138)
 52. Veness J, Silver D, Uther WT, Blair A. 2009 Bootstrapping from game tree search. In *NIPS*, vol. 19 (eds Y Bengio, D Schuurmans, J Lafferty, C Williams, A Culotta), pp. 1937–1945. Red Hook, NY: Curran Associates.
 53. Dayan P. 1993 Improving generalization for temporal difference learning: the successor representation. *Neural Comput.* **5**, 613–624. (doi:10.1162/neco.1993.5.4.613)
 54. Sutton RS. 1995 TD models: modeling the world at a mixture of time scales. In *ICML*, vol. 12 (eds A Prieditis, SJ Russell), pp. 531–539. San Mateo, CA: Morgan Kaufmann.
 55. Silver D, Ciosek K. 2012 Compositional planning using optimal option models. In *Proc. 29th Int. Conf. on Machine Learning, ICML '12* (eds J Langford, J Pineau), pp. 1063–1070. New York, NY: Omni Press.
 56. Botvinick M, Weinstein A. 2014 Model-based hierarchical reinforcement learning and human action control. *Phil. Trans. R. Soc. B* **369**, 20130480. (doi:10.1098/rstb.2013.0480)
 57. Sutton RS, Precup D, Singh S. 1999 Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artif. Intell.* **112**, 181–211. (doi:10.1016/S0004-3702(99)00052-1)
 58. Ng AY, Parr R, Koller D. 1999 Policy search via density estimation. In *NIPS* (eds SA Solla, TK Leen, K Müller), pp. 1022–1028. Cambridge, MA: MIT Press.
 59. Wang T, Bowling M, Schuurmans D. 2007 Dual representations for dynamic programming and reinforcement learning. In *IEEE Int. Symp. on Approximate Dynamic Programming and Reinforcement Learning, Honolulu, HI, 1–5 April 2007. ADPRL 2007*, pp. 44–51. IEEE. (doi:10.1109/ADPRL.2007.368168)
 60. Dickinson A. 1985 Actions and habits: the development of behavioural autonomy. *Phil. Trans. R. Soc. Lond. B* **308**, 67–78. (doi:10.1098/rstb.1985.0010)
 61. Dearden R, Friedman N, Russell S. 1998 Bayesian Q-learning. In *Proc. Fifteenth Nat. Conf. on Artificial Intelligence*, pp. 761–768. Menlo Park, CA: American Association for Artificial Intelligence.
 62. Engel Y, Mannor S, Meir R. 2003 Bayes meets Bellman: the Gaussian process approach to temporal difference learning. In *ICML* (eds T Fawcett, N Mishra), pp. 154–161. Washington, DC: AAAI Press.
 63. Mannor S, Simester D, Sun P, Tsitsiklis JN. 2004 Bias and variance in value function estimation. In *Proc. 21st Int. Conf. on Machine Learning*, p. 72. New York, NY: Association for Computing Machinery.
 64. Moore AW, Atkeson CG. 1993 Prioritized sweeping: reinforcement learning with less data and less time. *Mach. Learn.* **13**, 103–130. (doi:10.1007/BF00993104)
 65. Adams CD. 1982 Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Q. J. Exp. Psychol.* **34**, 77–98.
 66. Watkins CJCH. 1989 Learning from delayed rewards. PhD thesis, Cambridge University, Cambridge, UK.
 67. Hay N, Russell SJ. 2011 Metareasoning for Monte Carlo tree search. Technical Report no. UCB/EECS-2011-119. EECS Department, University of California, Berkeley, CA.
 68. Bellman RE. 1957 *Dynamic programming*. Princeton, NJ: Princeton University Press.
 69. Balleine BW, O'Doherty JP. 2010 Human and rodent homologues in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48–69. (doi:10.1038/npp.2009.131)
 70. Devan BD, Hong NS, McDonald RJ. 2011 Parallel associative processing in the dorsal striatum: segregation of stimulus–response and cognitive control subregions. *Neurobiol. Learn. Mem.* **96**, 95–120. (doi:10.1016/j.nlm.2011.06.002)
 71. Thorn CA, Atallah H, Howe M, Graybiel AM. 2010 Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron* **66**, 781–795. (doi:10.1016/j.neuron.2010.04.036)
 72. Alexander GE, Crutcher MD. 1990 Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci.* **13**, 266–271. (doi:10.1016/0166-2236(90)90107-L)

73. Alexander GE, DeLong MR, Strick PL. 1986 Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* **9**, 357–381. (doi:10.1146/annurev.ne.09.030186.002041)
74. Frank MJ, Claus ED. 2006 Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, reversal. *Psychol. Rev.* **113**, 300–326. (doi:10.1037/0033-295X.113.2.300)
75. Balleine BW, Dickinson A. 1998 Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419. (doi:10.1016/S0028-3908(98)00033-1)
76. Corbit LH, Muir JL, Balleine BW. 2003 Lesions of mediadorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. *Eur. J. Neurosci.* **18**, 1286–1294. (doi:10.1046/j.1460-9568.2003.02833.x)
77. O'Doherty JP. 2007 Lights, camera, action! The role of human orbitofrontal cortex in encoding stimuli, rewards, choices. *Ann. NY Acad. Sci.* **1121**, 254–272. (doi:10.1196/annals.1401.036)
78. de Wit S, Watson P, Harsay HA, Cohen MX, van de Vijver I, Ridderinkhof KR. 2012 Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. *J. Neurosci.* **32**, 12 066–12 075. (doi:10.1523/JNEUROSCI.1088-12.2012)
79. Liljeholm M, Tricomi E, O'Doherty JP, Balleine BW. 2011 Neural correlates of instrumental contingency learning: differential effects of action–reward conjunction and disjunction. *J. Neurosci.* **31**, 2474–2480. (doi:10.1523/JNEUROSCI.3354-10.2011)
80. Dayan P, Berridge K. 2014 Model-based and model-free Pavlovian reward learning: revaluation, revision and revelation. *Cogn. Affect. Behav. Neurosci.* **14**, 473–492. (doi:10.3758/s13415-014-0277-8)
81. Prévost C, Liljeholm M, Tyszka JM, O'Doherty JP. 2012 Neural correlates of specific and general Pavlovian-to-instrumental transfer within human amygdala subregions: a high-resolution fMRI study. *J. Neurosci.* **32**, 8383–8390. (doi:10.1523/JNEUROSCI.6237-11.2012)
82. Colwill RM, Rescorla RA. 1988 Associations between the discriminative stimulus and the reinforcer in instrumental learning. *J. Exp. Psychol. Anim. Behav. Process.* **14**, 155–164. (doi:10.1037/0097-7403.14.2.155)
83. Estes W. 1943 Discriminative conditioning. I. A discriminative property of conditioned anticipation. *J. Exp. Psychol.* **32**, 150–155. (doi:10.1037/h0058316)
84. Hart G, Leung BK, Balleine BW. 2013 Dorsal and ventral streams: the distinct role of striatal subregions in the acquisition and performance of goal-directed actions. *Neurobiol. Learn. Mem.* **108**, 104–118. (doi:10.1016/j.nlm.2013.11.003)
85. Rescorla RA, Solomon RL. 1967 Two-process learning theory: relationships between Pavlovian conditioning and instrumental learning. *Psychol. Rev.* **74**, 151–182. (doi:10.1037/h0024475)
86. Corbit LH, Leung BK, Balleine BW. 2013 The role of the amygdala–striatal pathway in the acquisition and performance of goal-directed instrumental actions. *J. Neurosci.* **33**, 17 682–17 690. (doi:10.1523/JNEUROSCI.3271-13.2013)
87. Lingawi NW, Balleine BW. 2012 Amygdala central nucleus interacts with dorsolateral striatum to regulate the acquisition of habits. *J. Neurosci.* **32**, 1073–1081. (doi:10.1523/JNEUROSCI.4806-11.2012)
88. Wiecki TV, Frank MJ. 2010 Neurocomputational models of motor and cognitive deficits in Parkinson's disease. *Prog. Brain Res.* **183**, 275–297. (doi:10.1016/S0079-6123(10)83014-6)
89. Foster DJ, Wilson MA. 2006 Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683. (doi:10.1038/nature04587)
90. Foster DJ, Wilson MA. 2007 Hippocampal theta sequences. *Hippocampus* **17**, 1093–1099. (doi:10.1002/hipo.20345)
91. Johnson A, van der Meer MAA, Redish AD. 2007 Integrating hippocampus and striatum in decision-making. *Curr. Opin. Neurobiol.* **17**, 692–697. (doi:10.1016/j.conb.2008.01.003)
92. Pfeiffer BE, Foster DJ. 2013 Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79. (doi:10.1038/nature12112)
93. Jones MW, Wilson MA. 2005 Theta rhythms coordinate hippocampal–prefrontal interactions in a spatial memory task. *PLoS Biol.* **3**, e402. (doi:10.1371/journal.pbio.0030402)
94. Lansink CS, Goltstein PM, Lankelma JV, McNaughton BL, Pennartz CM. 2009 Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* **7**, e1000173. (doi:10.1371/journal.pbio.1000173)
95. Doya K, Ishii S, Pouget A, Rao RP. (eds) 2007 *Bayesian brain: probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.
96. Rainer G, Rao SC, Miller EK. 1999 Prospective coding for objects in primate prefrontal cortex. *J. Neurosci.* **19**, 5493–5505.
97. Sakai K, Miyashita Y. 1991 Neural organization for the long-term memory of paired associates. *Nature* **354**, 152–155. (doi:10.1038/354152a0)
98. Faure A, Haberland U, Condé F, El Massioui N. 2005 Lesion to the nigrostriatal dopamine system disrupts stimulus–response habit formation. *J. Neurosci.* **25**, 2771–2780. (doi:10.1523/JNEUROSCI.3894-04.2005)
99. Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. 2013 A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973. (doi:10.1038/nn.3413)
100. Wang LP, Li F, Wang D, Xie K, Wang D, Shen X, Tsien JZ. 2011 NMDA receptors in dopaminergic neurons are crucial for habit learning. *Neuron* **72**, 1055–1066. (doi:10.1016/j.neuron.2011.10.019)
101. Dickinson A, Smith J, Mirenovic J. 2000 Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behav. Neurosci.* **114**, 468–483. (doi:10.1037/0735-7044.114.3.468)
102. Robinson MJF, Berridge KC. 2013 Instant transformation of learned repulsion into motivational 'wanting'. *Curr. Biol.* **23**, 282–289. (doi:10.1016/j.cub.2013.01.016)
103. Botvinick MM, Cohen J. 2014 The computational and neural basis of cognitive control: charted territory and new frontiers. *Cogn. Sci.* (doi:10.1111/cogs.12126)
104. Frank MJ, Loughry B, O'Reilly RC. 2001 Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cogn. Affect. Behav. Neurosci.* **1**, 137–160. (doi:10.3758/CABN.1.2.137)
105. Hazy TE, Frank MJ, O'Reilly RC. 2006 Banishing the homunculus: making working memory work. *Neuroscience* **139**, 105–118. (doi:10.1016/j.neuroscience.2005.04.067)
106. Miller EK, Cohen JD. 2001 An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202. (doi:10.1146/annurev.neuro.24.1.167)
107. O'Reilly R, Braver T, Cohen J. 1999 A biologically based computational model of working memory. In *Models of working memory: mechanisms of active maintenance and executive control* (eds A Mikay, P Shah), pp. 375–411. New York, NY: Cambridge University Press.
108. O'Reilly RC, Frank MJ. 2006 Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* **18**, 283–328. (doi:10.1162/089976606775093909)
109. O'Reilly RC, Noelle DC, Braver TS, Cohen JD. 2002 Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cereb. Cortex* **12**, 246–257. (doi:10.1093/cercor/12.3.246)
110. Rougier NP, Noelle DC, Braver TS, Cohen JD, O'Reilly RC. 2005 Prefrontal cortex and flexible cognitive control: rules without symbols. *Proc. Natl Acad. Sci. USA* **102**, 7338–7343. (doi:10.1073/pnas.0502455102)
111. Howes A, Lewis RL, Vera A. 2009 Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychol. Rev.* **116**, 717–751. (doi:10.1037/a0017187)
112. Lewis RL, Howes A, Singh S. 2014 Computational rationality: linking mechanism and behavior through utility maximization. *Top. Cogn. Sci.* **6**, 279–311. (doi:10.1111/tops.12086)
113. Raby CR, Alexis DM, Dickinson A, Clayton NS. 2007 Planning for the future by western scrub-jays. *Nature* **445**, 919–921. (doi:10.1038/nature05575)
114. Loewenstein G, Prelec D. 1992 Anomalies in intertemporal choice: evidence and an interpretation. *Q. J. Econ.* **107**, 573–597. (doi:10.2307/2118482)
115. Guthrie E. 1935 *The psychology of learning*. New York, NY: Harper.
116. Botvinick MM, Niv Y, Barto AC. 2009 Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* **113**, 262–280. (doi:10.1016/j.cognition.2008.08.011)

117. Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, Botvinick MM. 2011 A neural signature of hierarchical reinforcement learning. *Neuron* **71**, 370–379. (doi:10.1016/j.neuron.2011.05.042)
118. Fermin A, Yoshida T, Ito M, Yoshimoto J, Doya K. 2010 Evidence for model-based action planning in a sequential finger movement task. *J. Motiv. Behav.* **42**, 371–379. (doi:10.1080/00222895.2010.526467)
119. Kool W, Botvinick M. 2013 The intrinsic cost of cognitive control. *Behav. Brain Sci.* **36**, 697–698. (doi:10.1017/S0140525X1300109X)
120. Shenhav A, Botvinick MM, Cohen JD. 2013 The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–240. (doi:10.1016/j.neuron.2013.07.007)
121. Michie D, Chambers R. 1968 Boxes: an experiment in adaptive control. *Mach. Intell.* **2**, 137–152.
122. Minsky M. 1952 *A neural-analogue calculator based upon a probability model of reinforcement*. Cambridge, MA: Harvard University, Psychological Laboratories.
123. Charness N. 1992 The impact of chess research on cognitive science. *Psychol. Res.* **54**, 4–9. (doi:10.1007/BF01359217)
124. de Groot AD. 1965 *Thought and choice in chess*. The Hague, The Netherlands: Mouton.
125. Newell A, Shaw JC, Simon HA. 1958 Chess-playing programs and the problem of complexity. *IBM J. Res. Dev.* **2**, 320–335. (doi:10.1147/rd.24.0320)
126. Newell A *et al.* 1972 *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
127. Anderson JR. 1993 Problem solving and learning. *Am. Psychol.* **48**, 35–44. (doi:10.1037/0003-066X.48.1.35)
128. Simon HA. 1975 The functional equivalence of problem solving skills. *Cogn. Psychol.* **7**, 268–288. (doi:10.1016/0010-0285(75)90012-2)
129. Huys QJ, Eshel N, O’Nions E, Sheridan L, Dayan P, Roiser JP. 2012 Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput. Biol.* **8**, e1002410. (doi:10.1371/journal.pcbi.1002410)