nature neuroscience

Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control

Nathaniel D Daw[1], Yael Niv[1,2] & Peter Dayan[1]

A broad range of neural and behavioral data suggests that the brain contains multiple systems for behavioral choice, including one associated with prefrontal cortex and another with dorsolateral striatum. However, such a surfeit of control raises an additional choice problem: how to arbitrate between the systems when they disagree. Here, we consider dual-action choice systems from a normative perspective, using the computational theory of reinforcement learning. We identify a key trade-off pitting computational simplicity against the flexible and statistically efficient use of experience. The trade-off is realized in a competition between the dorsolateral striatal and prefrontal systems. We suggest a Bayesian principle of arbitration between them according to uncertainty, so each controller is deployed when it should be most accurate. This provides a unifying account of a wealth of experimental evidence about the factors favoring dominance by either system.

Diverse neural systems, notably prefrontal cortex, the striatum and their dopaminergic afferents, are thought to contribute to the selection of actions. Their differential and integrative roles are under active examination, and an important hypothesis is that subparts of these regions subserve two largely distinct and parallel routes to action. Such a division is the neurobiological scaffolding for an equivalent hypothesis about dual controllers that is prominent in psychological accounts of a range of behavioral phenomena in economic, social and animal-conditioning contexts[1–5].

The conventional idea is that the dorsolateral striatum and its dopaminergic afferents support habitual or reflexive control[6], whereas prefrontal cortex is associated with more reflective or cognitive action planning[7]. (Following this convention, we will refer to the cognitive circuit as 'prefrontal', although it likely involves a number of additional regions, potentially including more medial striatal territories[8].) This suggested dissociation is consistent with a range of electrophysiological[9–11], functional magnetic resonance imaging (fMRI)[12,13] and lesion studies[14–17]. The last are based on a clever behavioral approach to differentiating dual control strategies: namely, conditioning studies in which the values of rewards are unexpectedly changed. Outcome re-valuation affects the two styles of control differently and allows investigation of the characteristics of each controller, its neural substrates and the circumstances under which it dominates.

Despite the wealth of evidence, there are few answers to two key normative questions: why should the brain use multiple action controllers, and how should action choice be determined when they disagree? For a framework for answers, we turn to reinforcement learning[18], the computational theory of learned optimal action control. In reinforcement learning, candidate actions are assessed through predictions of their values, defined in terms of the amount of reward they are expected eventually to bring about. Such predictions pose statistical and computational challenges when reward is contingent on performing a sequence of actions, and thus early action choices incur only deferred rewards. Approximations are essential in the face of these challenges; there are two major classes of reinforcement learning, which make different approximations, and so are differentially accurate in particular circumstances. One class involves 'model-free' approaches such as temporal-difference learning, which underpin existing popular accounts of the activity of dopamine neurons and their (notably dorsolateral) striatal projections[19,20]. The other class involves 'model-based' methods[18], which we identify with the second, prefrontal cortex system.

We propose that the difference in the accuracy profiles of different reinforcement learning methods both justifies the plurality of control and underpins arbitration. To make the best decisions, the brain should rely on a controller of each class in circumstances in which its predictions tend to be most accurate. Here we suggest how the brain might estimate this accuracy for the purpose of arbitration by tracking the relative uncertainty of the predictions made by each controller. We show that this accounts for a range of factors shown in behavioral studies to favor either controller. To isolate our hypothesis, we develop the bulk of our account assuming strict separation between the systems; other aspects of their integration, particularly through learning[21], are certainly also important.

We interpret the two controllers as representing opposite extremes in a trade-off between the statistically efficient use of experience and computational tractability. Temporal-difference learning[18] is a model-free reinforcement learning method, which offers a compelling account of the activity of dopamine neurons in classical and instrumental

[1]Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, UK. [2]Interdisciplinary Center for Neural Computation, Hebrew University. P.O. Box 1255, Jerusalem 91904, Israel. Correspondence should be addressed to N.D.D. (daw@gatsby.ucl.ac.uk).

Received 15 April; accepted 12 September; published online 6 November 2005; doi:10.1038/nn1560

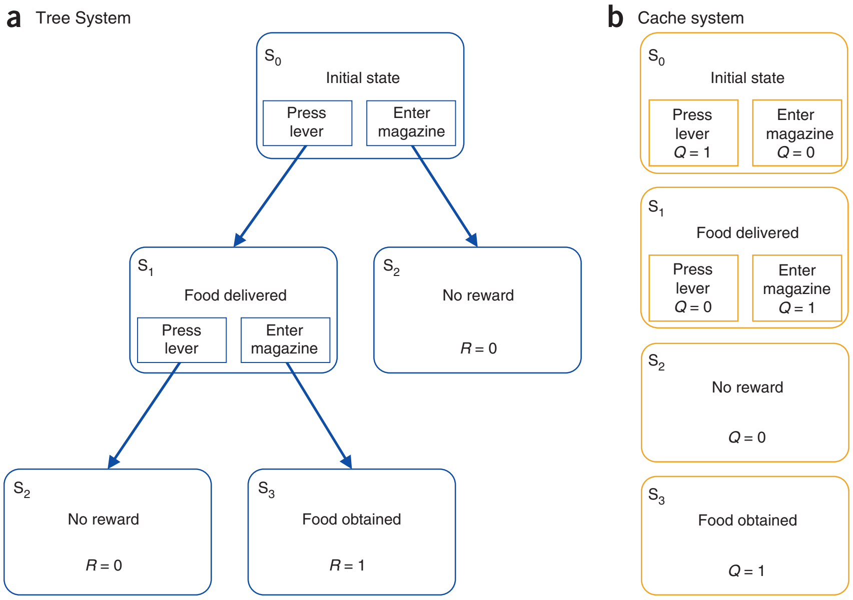


Figure 1 Task representations used by tree-search and caching reinforcement learning methods in a discrete-choice, discrete-trial representation of a standard instrumental conditioning task. (**a**) Structure of the task as represented by a tree-search controller. S_0–S_3 are the four possible states within the task; $R = \{1, 0\}$ represents whether or not reward was attained. (**b**) A caching reinforcement learning controller represents only the scalar expected future value ('Q') for each action in each state, divorced from the actual sequence and identity of future consequences.

learning tasks[19,20]. The foundation of this method is what we refer to as 'caching': namely, the association of an action or situation with a scalar summary of its long-run future value. A hallmark of this is the ubiquitous transfer of the dopaminergic response from rewards to the stimuli that predict them[20]. Working with cached values is computationally simple but comes at the cost of inflexibility: the values are divorced from the outcomes themselves and so do not immediately change with the re-valuation of the outcome. This is also the defining behavioral characteristic of habitual control.

By contrast, we suggest that the prefrontal circuit subserves a model-based reinforcement learning method. This constructs predictions of long-run outcomes, not through cached storage, but rather on the fly, by chaining together short-term predictions about the immediate consequences of each action in a sequence. Because this involves exploring a branching set of possible future situations, such methods are also known as 'tree search'. Search in deep trees can be expensive in terms of memory and time and can also be error-prone. However, that the predictions are constructed on the fly allows them to react more nimbly to changed circumstances, as when outcomes are re-valued. This, in turn, is the behavioral hallmark of cognitive (or 'goal-directed') control.

Here we develop these ideas in a formal, computational model and present simulation results that demonstrate the model's ability to capture a body of animal conditioning data concerning the trade-off between controllers. Our results suggest that principles of sound, approximate, statistical reasoning may explain why organisms use multiple decision-making strategies and also provide a solution to the problem of arbitrating between them.

RESULTS

Post-training reinforcer devaluation

We begin by discussing key experimental results suggesting the circumstances under which each controller dominates. Behavioral psychologists have investigated this issue extensively by post-training reinforcer devaluation (see a recent review[5] for references). In a typical experiment, hungry rats are trained to perform a sequence of actions, usually a lever press followed by entry to a food magazine, to obtain a reward such as a food pellet. We formally depict this task (**Fig. 1a**) as a tree of possible situations (states) that the subject can face in the task, the transitions between those states engendered by the possible actions and the reward that is available given an appropriate sequence of actions. Acquiring this arboreal representation of the task from experience, and using it to choose appropriate actions, are exactly the goals of the tree-search controller.

In the next phase of the experiment, the value of the food pellets is reduced, for instance by prefeeding the animal with them or by pairing them with illness to induce aversion. Then, animals are tested to see if they will continue to perform the actions previously associated with the newly devalued outcome. The test is performed without delivering outcomes (formally, in extinction) to prevent new learning about the value of the outcome during these trials.

Outcome devaluation exploits a key distinction between tree search (**Fig. 1a**) and caching (**Fig. 1b**). Only tree search enumerates the specific consequences expected for some course of action, such as the identity of the food reward expected. The cached value of an action is, by its nature, independent of any such specific outcome information. Thus, if an animal acts based on a cached value, it will continue to do so even after the outcome has been devalued. In psychology, such outcome-insensitive behavior is known as 'habitual'[5,22]. If, however, a behavior is determined by tree search, its propensity should be sharply reduced following devaluation. Psychologists term such outcome-sensitive behavior 'goal-directed'[5,22], as it changes when 'goals' are re-valued.

Behavioral experiments (summarized in **Fig. 2**) demonstrate that, under different circumstances, animals show both profiles of devaluation sensitivity. Moderately trained lever presses are indeed sensitive to outcome devaluation (**Fig. 2a**, left)—suggesting control by a tree-search system. After extensive training, though, lever pressing becomes insensitive to devaluation (**Fig. 2a**, middle)—suggesting a transition to caching control[23]. Lesions or depletions of dopaminergic input to dorsolateral areas of the striatum evidently block this transfer of control to a caching system[14,24]. Such animals display relatively normal learning of the task, but despite over-training, their lever pressing is persistently sensitive to devaluation. This is consistent with choice relying only on an intact tree-search controller.

The transition to caching with over-training is also tempered by two factors—the complexity of action choice and the proximity of the action to reward. In more complex tasks, in which an animal may, for instance, execute either of two different actions to obtain two different rewards, extensively trained actions remain sensitive to outcome devaluation (**Fig. 2a**, right), indicating a dominance of tree-search control[25,26]. Finally, though the evidence is perhaps less persuasive, actions closer to the reward are more sensitive to devaluation than

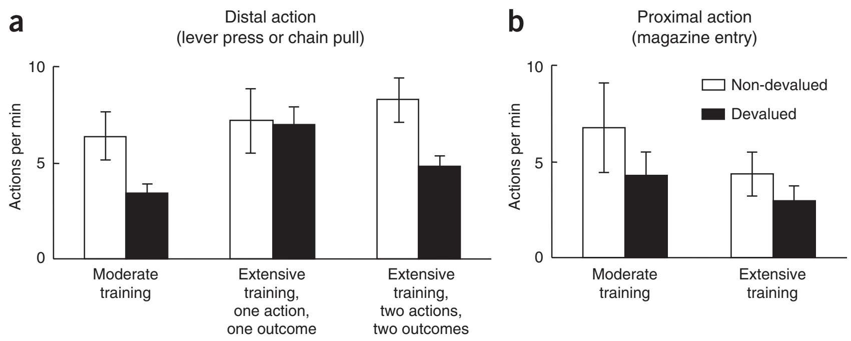


Figure 2 Behavioral results from reward devaluation experiments in rats. Actions per minute in an extinction test after devaluation of the outcome (black) or without devaluation (white). (**a**) Actions distal from the outcome (lever pressing and chain pulling) after moderate or extensive training and with one or two actions and outcomes, adapted from ref. 26, experiment 2. (**b**) Magazine entries (more proximal to the outcome), adapted from ref. 17. Data and error bars reproduced here are for a control group; differences were significant when collapsed with two additional lesion groups. Error bars: s.e.m.

actions further away. For instance, when animals must press a lever and then enter a food magazine to obtain reward, the action more proximal to reward—magazine entry—remains devaluation-sensitive even after extensive training[17] (**Fig. 2b**, right). In the experiment depicted here, this effect was significant only when collapsed over multiple lesion groups (which did not differ significantly among themselves), and there are only few other published reports of over-trained magazine behavior. However, actions more proximal to reward are more readily sensitive to devaluation in 'incentive learning' studies[27], and extensively trained magazine responses remain devaluation-sensitive in a Pavlovian task without lever pressing[28].

The counterpart to lesions affecting the caching system is that lesions to a wide network of structures—including prelimbic cortex (a subarea of rat prefrontal cortex)[15–17], prefrontal-associated regions of dorsomedial striatum[8], basolateral amygdala[29], gustatory insular cortex[30] and, in a monkey study, orbitofrontal cortex[31]—seem to interfere with tree-search control. That is, they eliminate devaluation sensitivity even for moderately trained behaviors.

Theory sketch

The lesion studies indicate that each controller can substitute for the other even under circumstances when it would not normally dominate. This suggests a theory combining separate and parallel reinforcement learners (the implementation is detailed in Methods and **Supplementary Methods** online).

As in previous applications of reinforcement learning to neural and behavioral data[20], we work with a stylized description of the experimental tasks (**Figs. 1a** and **3**). This allows us to expose a unifying, normative interpretation of the pattern of experimental results discussed above, rather than focusing on a quantitative fit to rather qualitative data. Here, the goal of optimal control is to choose actions that maximize the probability of the ultimate receipt of a valued outcome (although it would be straightforward to include additional factors into the optimization, such as risk-sensitivity in the case of stochastic rewards). Optimization can be accomplished by calculating or learning the value of taking each action at each state, defined in terms of the probability that reward will later be received when starting from that action in that state. Given such information, advantageous actions can be chosen simply by comparing their values. The collection of values is called a 'state-action value function' or, for simplicity, a value function.

The two classes of reinforcement learning methods can produce different, and differentially accurate, estimates of the value function. As in other cases of evidence reconciliation in neuroscience, such as multisensory integration[32], we suggest that arbitration between values is based on the uncertainty or expected inaccuracy of each. Uncertainty quantifies ignorance about the true values (for example, about the probabilities of different payoffs); it should be distinguished from risk (which generically arises when payoffs are stochastic, but their probabilities may be known). For simplicity, we assume that the estimated value of each action is taken to be that derived from the controller that is more certain about the value. (Though most reliable, this estimate is not necessarily the largest.) The probability of choosing an action for execution is then proportional to this value. In addition to controlling estimation, uncertainty about an action's value might, in principle, influence choice directly, as by promoting exploration to seek undiscovered rewards.

In general, both controllers are uncertain about the values because they begin ignorant and have only a limited amount of noisy experience. Even given infinite experience, uncertainty persists, due to the possibility that the task itself (and hence the long-term values) can

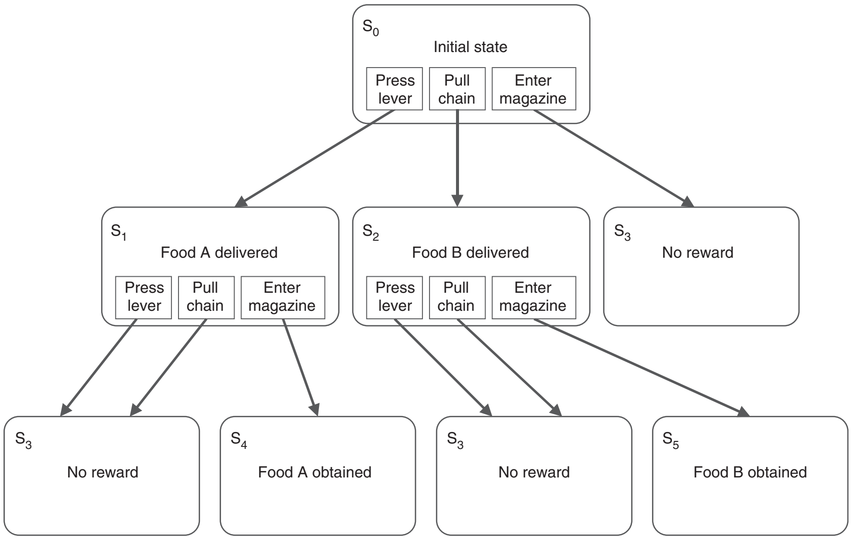


Figure 3 Stylized tree representation of an instrumental conditioning task with two actions (a lever press and a chain pull) for two rewards. Note the additional states compared with **Figure 1a**.

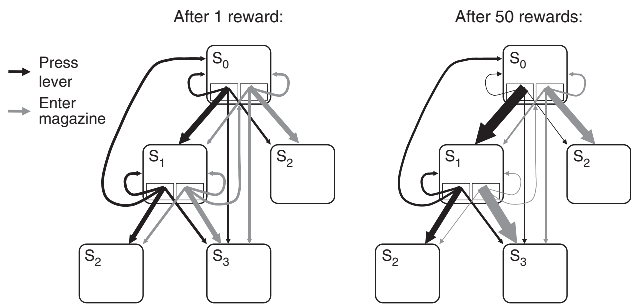


Figure 4 Tree estimation at two stages of learning by the tree-search system on the task of **Figure 1a**. States are as in that figure; arrows represent the state transitions expected following a lever press or a magazine entry, with the width of each proportional to the estimated probability of the transition (averaged over 250 runs; s.e.m. error bars negligible). Before learning, all transitions were equally likely (data not shown).

change unexpectedly. We quantify uncertainty using approximate Bayesian versions of each reinforcement learning algorithm[33,34]. The differing methods of value estimation of the two systems give rise to differing uncertainty profiles.

A (prefrontal) tree-search system uses experience with the task to estimate the nature of the state transitions and rewards (essentially, reconstructing the 'trees' of **Figs. 1a** and **3**). Long-term reward probabilities are estimated by iteratively searching through this tree; uncertainty about which tree describes the task makes the value estimates uncertain. Furthermore, such tree search is computationally demanding in realistic (wide or deep) trees. Thus, in practice, approximations must be introduced at each iteration, such as 'pruning' or exploring only a subset of paths. We model the resulting inaccuracy or 'computational noise' in derived value estimates as an additional source of uncertainty that accumulates with each search step.

A (dorsolateral striatal) caching system such as temporal-difference learning[18] estimates the long-term values directly from experience, without explicitly constructing a tree. This relies on a different approximation: 'bootstrapping', or using the value estimates cached for subsequently encountered states as stand-ins for the actual long-term values at predecessor states. Initial ignorance and continual change make these values potentially inaccurate, and thus cause uncertainty. By contrast, though, the cached values make calculation straightforward, so there is little computational 'noise' associated with its output.

To summarize, both the value predictions and the estimated uncertainties will differ between the tree-search and caching systems. Our account of action choice is based on an analysis of these two sets of quantities.

Simulations

We simulated the two-controller reinforcement learning model on the action choice tasks, beginning with the task with one lever press for one outcome (**Fig. 1a**). The quantitative results conformed with the qualitative expectations adduced above. The prefrontal tree system learned, over experience, the structure of action-induced state transitions in the task, assigning high probability to the actual transitions (**Fig. 4**; the system additionally tracked uncertainty about its estimates of the transition probabilities, which is not illustrated).

We studied each system's action values (here, posterior means), along with its uncertainty (posterior variances) about those values, as functions of the amount of training and the position of the action in the behavioral sequence relative to the reward (**Fig. 5a**,**b**). Each system's prior ignorance gradually resolved with experience. In all simulations, model-based reinforcement learning was more confident early in training, even though both systems had matched initial uncertainty. This is because under prefrontal tree search, any particular morsel of experience immediately propagates to influence the estimates of action values at all states; the effect of bootstrapping in dorsolateral striatal temporal-difference learning is to delay such propagation, making the system less data-efficient.

Because the systems incorporate the expectation that actions' values may change, past observations gradually become less relevant to present value estimates. This effective time horizon on the data implies that the uncertainties asymptote at finite levels for both systems. For the same reason, the value predictions can asymptote well short of the true payoffs. This asymptotic uncertainty has a harsher effect on the data-inefficient cache. Thus, for the action proximal to reward (the magazine response), enhanced data efficiency allowed the tree-search system to be more certain, even asymptotically (**Fig. 5b**). However, an extra iteration of tree search was required to evaluate the action more distal from reward (the lever press), incurring additional uncertainty asymptotically due to the assumption of computational noise outlined above. The effect of this, asymptotically (**Fig. 5a**), was to favor the cache system, which suffers no such computational noise because it recalls values rather than computing them.

We saw different results for a version of the task with two actions for two outcomes (**Fig. 6a**,**b**). Here, the agent's experience was spread between more states and actions. Given the expectation of task change and the resulting temporal horizon on past experience, fewer relevant

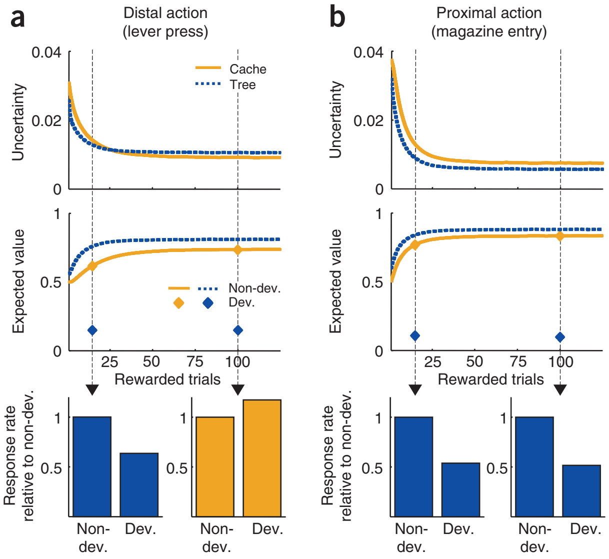


Figure 5 Simulation of the dual-controller reinforcement learning model in the task of **Figure 1a**. (**a**) Distal action (lever press); (**b**) Proximal action (magazine entry). The topmost graphs show uncertainties (posterior variances) in the value estimates for different actions according to the cache (blue line) and tree (gold line), as a function of the number of rewarded training trials. The middle graphs show the value estimates themselves (posterior means); diamonds indicate the value estimates that would result after reward devaluation at various stages of training. Beneath the graphs are bar plots comparing the probability of choosing actions before and after their consequences were devalued, normalized to the non-devalued level. Bar color denotes which system (cache: blue; tree: gold) controlled the action in a majority of the 250 runs. All data reported are means over 250 runs; error bars (s.e.m.) are negligible.

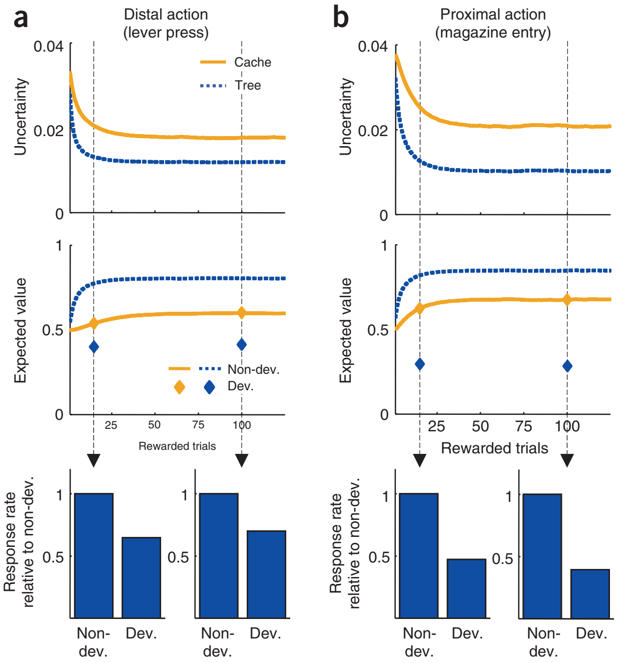


Figure 6 Simulation of the dual-controller reinforcement learning model in the task of **Figure 3**, in which two different actions produced two different rewards. One of the rewards was devalued in probe trials. (**a**) Distal action (lever press). (**b**) Proximal action (magazine entry). The same conventions and results are shown as in **Figure 5**, except that data reported are means over 1,000 runs; error bars (s.e.m.) are again negligible.

data were available to constrain any particular action value. The effect was asymptotically to preserve the tree system's uncertainty advantage from the early training, low-data situation, even for the distal lever press (**Fig. 6a**).

Whenever the tree system dominated, the overall system's action choices were sensitive to outcome devaluation, whereas when the caching system dominated, they were not (**Figs. 5** and **6**, bar plots). This is because the underlying value predictions were sensitive to devaluation only in the tree system. The simulations, then, reproduced and explained the pattern seen in the behavioral experiments: overtraining promoted devaluation insensitivity, unless opposed by the countervailing effects of proximity to reward or task complexity. The results support the underlying hypothesis that the brain appropriately deploys each controller under those circumstances in which it is expected to be most accurate.

DISCUSSION

Our account builds on and extends existing ideas in several key ways. In contrast to the somewhat descriptive animal learning theories that are its foundation[4,5,22], we have adopted a normative view, unifying the body of results on controller competition by appealing to uncertainty. This stance also contrasts with accounts of human behavioral data[1,3]: notably, ideas in economics[2] suggesting that irrational, impulsive or emotional limbic influences (in our terms, the caching system) interfere with a more rational prefrontal controller. Under our account, both controllers are pursuing identical rational ends; in appropriate circumstances, the caching controller can more effectively accomplish the same functions as the prefrontal controller.

Among reinforcement learning theories, there are various precedents for the idea of combining several controllers, including multiple caching controllers[35–37] and also (partly cerebellar) model-based and model-free controllers[38]. However, normative, competitive interaction has not hitherto been investigated. Most other reinforcement learning theories that contemplate model-based control either completely replace caching with search[39,40], or envision a hybrid blending features of both[41,42]. Such theories founder on lesion results indicating a dissociation between the neural substrates for tree-like and cache-like choice[8,14–17,24].

Of course, normativity only extends so far for us. The true reason for multiple controllers in our theory is the computational intractability of the complete Bayesian solution (roughly speaking, the tree-search system unencumbered by computational incapacity) and the resulting need for approximations. The cache system is an extreme example of an approximation that embraces potential inaccuracy to gain computational simplicity.

Neural substrates

We built on the classic idea that habitual control is associated with dopamine and dorsolateral striatum, and more cognitive search with prefrontal cortex. Because behavioral and lesion studies suggest these controllers can operate independently, for the purpose of modeling we made the simplifying approximation that they are strictly separate. However, their neural substrates are clearly intertwined—prefrontal cortex is itself dopaminergically innervated, and cortex and striatum are famously interconnected in 'loops'[43], including one that joins prefrontal areas with dorsomedial subregions of striatum. Indeed, recent pharmacological and lesion results implicate those prefrontal-associated striatal areas in tree-search control[8]. Competition between model-based and model-free control might, therefore, best be viewed as between dorsomedial and dorsolateral corticostriatal loops, rather than between cortex and striatum per se, a view that extends previous ideas about multiple caching controllers coexisting in different loops[35,36]. Although dopamine is hypothesized to support learning in the caching system, the role of dopamine in the tree-search controller remains wholly unresolved.

Computational considerations also suggest that the systems should interact. Owing to computational costs in tasks involving deep trees, it is commonplace in reinforcement learning to search partway along some paths, then use cached values to substitute for unexplored subtrees[18]. Uncertainties can be compared at each stage to decide whether to expand the tree or to fall back on the cache[44], trading off the likely costs (for example, time or calories) of additional search against its expected benefits (more accurate valuations allowing better reward harvesting). The essentials of our account would be preserved in a model incorporating such partial evaluation, and the resulting improvement in the tree system's valuations due to learning in the cache system echoes other suggestions that learning in the basal ganglia might train or inform cortex[11,21].

There is limited evidence about the substrate for the uncertainty-based arbitration that has been our key focus. First, along with direct, population-code representations of uncertainty[45], cholinergic and noradrenergic neuromodulation have often been implicated[46]. Second, two candidates for arbitration are the infralimbic cortex (IL; part of the prefrontal cortex) and the anterior cingulate cortex (ACC). Lesions to the IL reinstate tree-search from previously caching control[16,17]; however, because this area is not classically part of the habitual system, it has been suggested that it might support controller competition[17]. The involvement of the ACC in the competition-related functions of monitoring and resolving response error and conflict has been suggested in experimental and theoretical studies[47,48].

Complementary evidence about dual control arises from spatial tasks in both humans and animals[37,49]. Navigational decisions can arise from a flexible 'cognitive map' that supports latent learning and is associated with the hippocampus; with practice, they become habitized and evidently under dorsal striatal control.

Experimental considerations

One route to test our framework is neuronal recordings. We expect activity in areas associated with each controller to reflect its decision preferences, even when (as a result of arbitration) the other is actually directing behavior. Behavior should thus be better correlated with activity in whichever system is producing it. By manipulating factors such as the amount of training or the proximity of response to reward, it should be possible to transfer control between the systems and thereby to switch the behavioral-neural correlations.

Researchers have recently recorded from striatum and prefrontal cortex (interpreted as parts of the cache and tree systems, respectively) in monkeys over-trained on an associative learning task with reversals[11]. Various features of this task could promote the dominance of either system—extreme over-training and task simplicity favor cache control, but action-reward proximity and frequent reversals promote tree search. The neural recordings are also inconclusive. A direct interpretation supports striatal control: neurons there are more strongly selective for the animal's choices, earlier in trials, and more robustly after reversals. However, an alternative interpretation instead supports prefrontal dominance, because change in the prefrontal representation correlates with behavioral re-learning following reversal. A devaluation challenge or recordings under different task circumstances (over-training levels, etc.) could help to distinguish these possibilities.

Because, in these recordings, representational changes occur faster in striatum, the authors suggest[11] that relearning the correct responses following reversal might be more rapid in the striatum, and that this knowledge subsequently transfers to cortex[21]. This contrasts with some habitization models in which learning progresses in the other order, though our theory makes no specific claim about the relative 'learning rates' of the two systems. In any case, subsequent analysis of error trials shows that the striatal firing reflects the animal's actual (and not the correct) choices (A. Pasupathy & E.K. Miller, *Comput. Syst. Neurosci. Abstr.*, p. 38, 2005). Finally, because the striatal region recorded (caudate) includes areas likely corresponding to dorsomedial striatum in rats, it may be that this area too is part of the tree system and not the cache[8], in which case properly interpreting the results will require a more finely fractionated understanding of the neural organization of tree search.

Our theory provides additional testable factors likely to influence the trade-off between systems. Computational pressures might be increased, and tree search discouraged, in tasks that pose more strenuous cognitive demands (for example, delayed match to sample; such a strategy has been used with humans[2] but not in animal devaluation studies). Introducing unexpected changes in task contingencies should also favor the data-efficient tree system, because relevant data thereby become more scarce. Further, although task complexity favors goal-directed control, the details of task structure may have subtler effects. It has long been known in reinforcement learning that caching is relatively advantageous in tasks with a fan-out structure (in which a state might be followed randomly by any of several others); conversely, tasks with linear or fan-in structure (several states leading to one) should favor search.

Finally, our theory is applicable to several other phenomena in animal behavior. Stimuli can signal reinforcement that is available irrespective of the animal's actions, and these 'Pavlovian' associations can affect behavior. Such stimulus-reward predictions might originate from both cache and tree systems, with rich interactions and consequences. In 'conditioned reinforcement', animals learn to work to receive a stimulus that had previously been paired with reinforcement. Such learning might occur in either of our reinforcement learning systems. However, it is also a plausible test case for their potential interaction through partial evaluation, as the tree system might explore the consequences of the (new) response but defer to the cache's evaluation of the (familiar) subsequent stimulus. Animals can acquire a new conditioned response even for a stimulus whose associated reinforcer had been devalued[50], suggesting at least the value of the stimulus was cached. The hypothesized involvement of both systems might be investigated with lesions disabling each.

Our theory also casts the phenomenon of 'incentive learning'[27] in a new light. In this, for some actions to be sensitive to outcome devaluation, the animal must previously have experienced the reinforcer in the devalued state. The predominant account of incentive learning[5] holds that such experience is necessary for the goal-directed system (our tree) to learn about the new value of the reinforcer. We suggest instead that experience with the outcome decreases the tree system's uncertainties (by confirming existing knowledge about the outcome's value). This tends to promote its dominance over the cache, explaining interactions between outcome exposure and other factors such as over-training and reward proximity[27]. Because outcome exposure allows the tree system to overcome caching control, our theory makes the strong prediction (contrary to the standard account) that the need for such experience should vanish in animals with lesions disabling the caching system.

METHODS

Background. For simplicity, we modeled conditioning tasks using absorbing Markov decision processes (MDPs)[18] (**Figs. 1a** and **3**)—ones in which experience is structured as a set of trials, with a set of terminal states at which an episode can end. We assumed that outcomes were delivered only (if at all) in terminal states and identified particular terminal states with particular outcomes (for instance, different foods).

Key to our account are two complicating factors. First, the agent started without knowing the exact MDP, which, furthermore, could change over time. These were the major sources of uncertainty. Second, although MDPs traditionally treat rewards with static, scalar utilities, here devaluation treatments explicitly changed some outcomes' utilities. For convenience, we assumed that rewards were binary (0 or 1) and used the probability that the reward was 1 in a particular terminal state as a surrogate for the associated reward's utility.

Choice in both cache and tree systems depended on scalar values—predictions of the future utility of executing a particular action at a particular state. If an outcome was devalued, both could learn by experiencing it that its corresponding state had lower utility. However, only the tree system used that information to guide subsequent action choice at distal states, as it derived action values by considering what future states would result. The cache system's values were stored scalars and were thus insensitive even to known changes in outcome value, absent new experience of the action actually producing the outcome.

Fully optimal choice in unknown MDPs is radically computationally intractable. Tree and cache reinforcement learning methods therefore each rely on approximations, and we tracked uncertainties about the values produced by such systems to determine for what circumstances each method is best suited.

Formal model. An absorbing MDP comprises sets $\mathcal{S}$ of states and $\mathcal{A}$ of actions, a 'transition function' $T(s, a, s') \equiv P(s(t+1) = s' \mid s(t) = s, a(t) = a)$ specifying the probability that state $s' \in \mathcal{S}$ will follow state $s \in \mathcal{S}$ given action $a \in \mathcal{A}$, and (in our version) a 'reward function' $R(s) \equiv P(\text{reward}(t) = 1 \mid s(t) = s)$ specifying the probability that reward is received in terminal state s.

Here, the state-action value function $Q(s, a)$ is the expected probability that reward will ultimately be received, given that the agent takes action a in state s

and chooses optimally thereafter. The formal definition is recursive:

$$Q(s,a) \equiv \begin{cases} R(s) & s \text{ is terminal } (a = \oslash) \\ \sum_{s'} T(s,a,s') \cdot \max_{a'}[Q(s',a')] & \text{otherwise} \end{cases}$$

Standard reinforcement learning methods[18] do not track uncertainty in their estimates of Q. We consider Bayesian variations[33,34], which estimate not simply the expected value $Q(s, a)$ but a posterior distribution $\mathbf{Q}_{s,a}(q) \equiv P(Q(s, a) = q \mid \text{data})$ that measures, for any $0 \le q \le 1$, how likely it is that the true optimal probability of future reward (compounded over different paths through the states) equals q, given the evidence, 'data', about transitions and outcomes so far observed. A Bayesian tree-search ('value iteration') system[34] uses experience to estimate a posterior distribution over the MDP (functions T and R) and explores it to derive distributions over $Q(s, a)$ (**Supplementary Fig. 1** online). A Bayesian caching ('Q-learning') system[33] instead stores a distribution over $Q(s, a)$ for each action and state and updates it for consistency with the stored value distributions of subsequently encountered states (**Supplementary Fig. 2** online). Full equations appear in **Supplementary Methods**.

If, for a particular controller, state and action, the distribution $\mathbf{Q}_{\text{s,a}}$ is sharply peaked at some q, then the controller is fairly certain of the value; if it is instead spread out over a range of possible q's, then the controller cannot identify the value with certainty. We thus arbitrated between the controllers' estimates on the basis of their variance (mean squared error, 'uncertainty'): given distributions $\mathbf{Q}_{s,a}^{\text{tree}}$ from the tree and $\mathbf{Q}_{s,a}^{\text{cache}}$ from the cache, we took the winning value $Q(s, a)$ to be the mean $\left\langle \mathbf{Q}_{s,a}^{\text{tree}} \right\rangle$ if the variance of $\mathbf{Q}_{s,a}^{\text{tree}}$ was smaller than the variance of $\mathbf{Q}_{s,a}^{\text{cache}}$, and the mean $\left\langle \mathbf{Q}_{s,a}^{\text{cache}} \right\rangle$ otherwise. (Softer integration schemes, such as a certainty-weighted average, are also possible.) Given winning estimates $Q(s, a)$ for each action available in the current state, we chose an action stochastically using softmax probabilities, $P(a(t) = a \mid s(t) = s) \propto e^{\beta Q(s,a)}$ where the parameter β controlled the tendency of the system to choose exclusively the action deemed best. Experimentally, the effect of devaluation can be assessed either within or between animals (by comparing to another action or group for which the outcome was not devalued). In our simulations, we compared the probabilities of choosing the same action a in the relevant state s, with or without devaluation (similar to the between-group approach); softmax action selection ensured that a reduction in $Q(s, a)$ for an action will reduce the probability that the action is chosen.

Note that posterior uncertainty quantifies ignorance about the true probability of reward, not inherent stochasticity in reward delivery. For instance, reward may follow from some state randomly with 50% probability—but if the controller can precisely identify that the true probability is 50% rather than some other number, the value is not uncertain.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS
We are grateful to B. Balleine, A. Courville, A. Dickinson, P. Holland, D. Joel, S. McClure and M. Sahani for discussions. The authors are supported by the Gatsby Foundation, the EU Bayesian Inspired Brain and Artefacts (BIBA) project (P.D., N.D.), a Royal Society USA Research Fellowship (N.D.) and a Dan David Fellowship (Y.N.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Kahneman, D. & Frederick, S. Representativeness revisited: attribute substitution in intuitive judgment. in *Heuristics and Biases: the Psychology of Intuitive Judgment* (eds. T. Gilovich, D.G. & Kahneman, D.) 49–81 (Cambridge University Press, New York, 2002).
2. Loewenstein, G. & O'Donoghue, T. Animal spirits: affective and deliberative processes in economic behavior. Working Paper 04–14, Center for Analytic Economics, Cornell University (2004).
3. Lieberman, M.D. Reflective and reflexive judgment processes: a social cognitive neuroscience approach. in *Social Judgments: Implicit and Explicit Processes* (eds. Forgas, J., Williams, K. & von Hippel, W.) 44–67 (Cambridge University Press, New York, 2003).
4. Killcross, S. & Blundell, P. Associative representations of emotionally significant outcomes. in *Emotional Cognition: from Brain to Behaviour* (eds. Moore, S. & Oaksford, M.) 35–73 (John Benjamins, Amsterdam, 2002).
5. Dickinson, A. & Balleine, B. The role of learning in motivation. in *Stevens' Handbook of Experimental Psychology Vol. 3: Learning, Motivation and Emotion* 3rd edn. (ed. Gallistel, C.R.) 497–533 (Wiley, New York, 2002).
6. Packard, M.G. & Knowlton, B.J. Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* **25**, 563–593 (2002).
7. Owen, A.M. Cognitive planning in humans: neuropsychological, neuroanatomical and neuropharmacological perspectives. *Prog. Neurobiol.* **53**, 431–450 (1997).
8. Yin, H.H., Ostlund, S.B., Knowlton, B.J. & Balleine, B.W. The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* **22**, 513–523 (2005).
9. Jog, M.S., Kubota, Y., Connolly, C.I., Hillegaart, V. & Graybiel, A.M. Building neural representations of habits. *Science* **286**, 1745–1749 (1999).
10. Holland, P.C. & Gallagher, M. Amygdala-frontal interactions and reward expectancy. *Curr. Opin. Neurobiol.* **14**, 148–155 (2004).
11. Pasupathy, A. & Miller, E.K. Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* **433**, 873–876 (2005).
12. McClure, S.M., Laibson, D.I., Loewenstein, G. & Cohen, J.D. Separate neural systems value immediate and delayed monetary rewards. *Science* **306**, 503–507 (2004).
13. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
14. Yin, H.H., Knowlton, B.J. & Balleine, B.W. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189 (2004).
15. Balleine, B.W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).
16. Coutureau, E. & Killcross, S. Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behav. Brain Res.* **146**, 167–174 (2003).
17. Killcross, S. & Coutureau, E. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex* **13**, 400–408 (2003).
18. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: an Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
19. Houk, J.C., Adams, J.L. & Barto, A.G. A model of how the basal ganglia generate and use neural signals that predict reinforcement. in *Models of Information Processing in the Basal Ganglia* (eds. Houk, J.C., Davis, J.L. & Beiser, D.G.) 249–270 (MIT Press, Cambridge, Massachusetts, 1995).
20. Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
21. Houk, J.C. & Wise, S.P. Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. *Cereb. Cortex* **5**, 95–110 (1995).
22. Dickinson, A. Actions and habits—the development of behavioural autonomy. *Phil. Trans. R. Soc. Lond. B* **308**, 67–78 (1985).
23. Adams, C.D. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Q. J. Exp. Psychol.* **34B**, 77–98 (1982).
24. Faure, A., Haberland, U., Condé, F. & Massioui, N.E. Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *J. Neurosci.* **25**, 2771–2780 (2005).
25. Colwill, R.M. & Rescorla, R.A. Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *J. Exp. Psychol. Anim. Behav. Process.* **11**, 520–536 (1985).
26. Holland, P.C. Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *J. Exp. Psychol. Anim. Behav. Process.* **30**, 104–117 (2004).
27. Balleine, B.W., Garner, C., Gonzalez, F. & Dickinson, A. Motivational control of heterogeneous instrumental chains. *J. Exp. Psychol. Anim. Behav. Process.* **21**, 203–217 (1995).
28. Holland, P. Amount of training affects associatively-activated event representation. *Neuropharmacology* **37**, 461–469 (1998).
29. Blundell, P., Hall, G. & Killcross, S. Preserved sensitivity to outcome value after lesions of the basolateral amygdala. *J. Neurosci.* **23**, 7702–7709 (2003).
30. Balleine, B.W. & Dickinson, A. The effect of lesions of the insular cortex on instrumental conditioning: evidence for a role in incentive memory. *J. Neurosci.* **20**, 8954–8964 (2000).
31. Izquierdo, A., Suda, R.K. & Murray, E.A. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *J. Neurosci.* **24**, 7540–7548 (2004).
32. Deneve, S. & Pouget, A. Bayesian multisensory integration and cross-modal spatial links. *J. Physiol. (Paris)* **98**, 249–258 (2004).
33. Dearden, R., Friedman, N. & Russell, S.J. Bayesian Q-learning. in *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI)* 761–768 (1998).
34. Mannor, S., Simester, D., Sun, P. & Tsitsiklis, J.N. Bias and variance in value function estimation. in *Proceedings of the 21st International Conference on Machine Learning (ICML)* 568–575 (2004).
35. Nakahara, H., Doya, K. & Hikosaka, O. Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *J. Cogn. Neurosci.* **13**, 626–647 (2001).
36. Tanaka, S.C. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.* **7**, 887–893 (2004).

37. Chavarriaga, R., Strosslin, T., Sheynikhovich, D. & Gerstner, W. A computational model of parallel navigation systems in rodents. *Neuroinformatics* **3**, 223–242 (2005).
38. Doya, K. What are the computations in the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Netw.* **12**, 961–974 (1999).
39. Suri, R.E. Anticipatory responses of dopamine neurons and cortical neurons reproduced by internal model. *Exp. Brain Res.* **140**, 234–240 (2001).
40. Smith, A.J., Becker, S. & Kapur, S. A computational model of the functional role of the ventral-striatal D2 receptor in the expression of previously acquired behaviors. *Neural Comput.* **17**, 361–395 (2005).
41. Dayan, P. & Balleine, B.W. Reward, motivation and reinforcement learning. *Neuron* **36**, 285–298 (2002).
42. Daw, N.D., Courville, A.C. & Touretzky, D.S. Timing and partial observability in the dopamine system. in *Advances in Neural Information Processing Systems 15*, 99–106 (MIT Press, Cambridge, Massachusetts, 2003).
43. Alexander, G.E., Delong, M.R. & Strick, P.L. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* **9**, 357–381 (1986).
44. Baum, E.B. & Smith, W.D. A Bayesian approach to relevance in game playing. *Artificial Intelligence* **97**, 195–242 (1997).
45. Pouget, A., Dayan, P. & Zemel, R.S. Inference and computation with population codes. *Annu. Rev. Neurosci.* **26**, 381–410 (2003).
46. Yu, A.J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).
47. Holroyd, C.B. & Coles, M.G. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* **109**, 679–709 (2002).
48. Botvinick, M.M., Cohen, J.D. & Carter, C.S. Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* **8**, 539–546 (2004).
49. Hartley, T. & Burgess, N. Complementary memory systems: competition, cooperation and compensation. *Trends Neurosci.* **28**, 169–170 (2005).
50. Parkinson, J.A., Roberts, A.C., Everitt, B.J. & Di Ciano, P. Acquisition of instrumental conditioned reinforcement is resistant to the devaluation of the unconditioned stimulus. *Q. J. Exp. Psychol. B* **58**, 19–30 (2005).

Supplementary methods

Here, we describe our implementation of uncertainty-tracking in the caching and tree-search systems, and then report the parameters governing these models and comment on the effect of slow changes in the task.

Recall that the goal of each system is to estimate, for each state s and action a, a (factorial) distribution $\mathbf{Q}_{s,a}$ over the future expected value $Q(s,a)$. The latter is defined:

$$Q(s,a) \equiv \begin{cases} R(s) & s \text{ is terminal } (a = \varnothing) \\ \sum_{s'} T(s,a,s') \cdot \max_{a'}[Q(s',a')] & \text{otherwise} \end{cases} \tag{1}$$

where $\varnothing$ stands in for the fact that no action is possible at a terminal state. Tree and cache systems use different approximations for value estimation, and we use the uncertainty as a measure of how appropriate these methods are in particular circumstances. However, the uncertainty computation itself requires a number of further shortcuts. In our implementation, these were matched between the systems so far as possible in the hope that even though the absolute values of the uncertainty measurements are likely erroneous, the errors should be similar between systems and thus their relative uncertainties should be informative about the actual reliabilities of the underlying value estimation methods.

Caching algorithm

We learned $\mathbf{Q}^{\text{cache}}$ from experience using "Bayesian Q-learning"[1] (for a different Bayesian formulation see Engel et al.[2]), adapted to our simplified class of MDPs. Notionally, the method involves assuming prior distributions $\mathbf{Q}^{\text{cache}}_{s,a}$ describing uncertainty about the value of each state and action, and then updating them using Bayes' theorem to reflect subsequent experience. We made four simplifying assumptions: that the distributions $\mathbf{Q}^{\text{cache}}_{s,a}$ were at each step expressed as beta distributions $Beta(\alpha_{s,a}, \beta_{s,a})$; that the distribution $\mathbf{Q}^{\text{cache}}_{s,a}$ was independent of $\mathbf{Q}^{\text{cache}}_{s',a'}$ for $s \neq s'$ or $a \neq a'$; that the distribution of the maximum (with respect to action choice) of a state-action value was just the distribution of values, $\mathbf{Q}^{\text{cache}}_{s,a}$, for the action a optimizing the mean $\langle \mathbf{Q}^{\text{cache}}_{s,a} \rangle$; and that the bootstrapped posterior distribution for a nonterminal state was specified by Dearden et al.'s[1] "mixture update" approximation. See Dearden et al.[1] for a full discussion of the merits of these simplifications; the most serious is the assumption that different states' values were independent, contrary to the coupling inherent in their definition (Equation 1). Our assumptions differ from Dearden's mainly in our use of a beta distribution for the posterior. We evaluated this last simplification against the same method implemented with an arbitrary posterior (finely discretized for numerical estimation), and found it innocuous. (In particular, over 500 steps of our task, the largest deviation between the methods' variance or mean estimates at any state was 0.5%). We nonetheless stress that the numerical accuracy of the approximate Bayesian computations is not key to our argument — we intended rather to implement both caching and tree-search learning using similar approximations, so that any computational biases impacted both similarly.

For the details of the method, consider being in state s — either s is a terminal state and we receive reward $r \in \{0, 1\}$, or we take action a and transition to another state s'. For terminal states s with prior $Beta(\alpha_{s,\varnothing}, \beta_{s,\varnothing})$ and reward r, Bayes' theorem specifies that the posterior value distribution $\mathbf{Q}^{\text{cache}}_{s,\varnothing}$ will be distributed as $Beta(\alpha_{s,\varnothing}{+}r, \beta_{s,\varnothing}{+}(1{-}r))$. For nonterminal states s followed by s', we wish to treat the successor state's mean value $\langle \mathbf{Q}^{\text{cache}}_{s',a'} \rangle$ (for the action a' optimizing that mean) as a bootstrapped sample of the predecessor state's mean value; the question is how to take into account uncertainty about the two states' values. If the successor state's value were $0 \leq q \leq 1$ we might, by analogy with the terminal state case, take the predecessor state's value posterior $\mathbf{Q}^{\text{cache}}_{s,a}$ as $Beta(\alpha_{s,a}{+}q, \beta_{s,a}{+}(1{-}q))$. Following Dearden et al.[1], we used the mixture of such distributions with respect to the successor state value distribution $\mathbf{Q}^{\text{cache}}_{s',a'}$:

$$\int_0^1 Beta(\alpha_{s,a}{+}q, \beta_{s,a}{+}(1{-}q)) \mathbf{Q}^{\text{cache}}_{s',a'}(q) dq \tag{2}$$

Though this integral is neither readily solvable nor itself a beta distribution, its mean and variance are analytic, and we thus approximated the predecessor state's posterior value $\mathbf{Q}^{\text{cache}}_{s,a}$ as the beta distribution matching those moments. They are:

$$\mu^{\text{cache}}_{s,a} = \frac{\alpha_{s,a} + \langle q \rangle^{\text{cache}}_{s',a'}}{n_{s,a} + 1} \tag{3}$$

$$(\sigma^2)^{\text{cache}}_{s,a} = \frac{1}{(n_{s,a}+2)(n_{s,a}+1)}(\alpha^2_{s,a} + \alpha_{s,a} + \langle q^2 \rangle^{\text{cache}}_{s',a'} + (2\alpha_{s,a}+1)\langle q \rangle^{\text{cache}}_{s',a'}) - (\mu^2)^{\text{cache}}_{s,a} \tag{4}$$

where $\langle q \rangle^{\text{cache}}_{s',a'}$ and $\langle q^2 \rangle^{\text{cache}}_{s',a'}$ are respectively the first and second moments of the beta distribution $\mathbf{Q}^{\text{cache}}_{s',a'}$ and $n_{s,a} = \alpha_{s,a} + \beta_{s,a}$.

To model outcome devaluation (e.g., treatments in which the animal is allowed to sample the outcome in the devalued state, but not in the context of the task), we replaced the distribution $\mathbf{Q}_{s,\varnothing}^{\text{cache}}$ for the terminal state s corresponding to the devalued outcome, reducing its expected value. Absent further learning with samples of trajectories ending in this state, this has no effect on the cached values of any other states.

Tree-search algorithm

A Bayesian tree-search method[3–5] involves two stages: model identification, and value computation. To identify the MDP, we assumed beta priors over the reward functions and, for each state and action, a Dirichlet prior over the vector of successor state probabilities. For these simulations, we assumed the number of states and their terminal or nonterminal status were known. As experience accrues, these distributions can be updated exactly by Bayes' theorem, simply by counting state transitions and rewards.

Given posterior distributions over the state transition and reward functions, a standard "certainty equivalent" method for estimating expected values $Q(s, a)$ would be to assume the true MDP is described by the means of those distributions and then to solve for the values using *value iteration*, or repeated application of Equation 1[6]. This is roughly equivalent to using tree search to compute the values of all states in parallel. Here we wish to quantify the uncertainty about the values that results from the uncertainty about the transition and reward functions. An optimal (if impractical) way to do so would be to repeat the value iteration process for all possible combinations of transition and reward functions, weighting each resulting set of state-action values by the probability of the transition and reward functions that produced it. Such an approach can be directly approximated by sampling from the distribution over MDPs[3,4]. Here, we used a set of approximations more closely matched to the Bayesian Q-learning methods discussed above — by performing not a set of iterations over future value for different trees, but rather a single iteration on the *distributions* over the future values implied by the distributions over the trees[5]. We assumed, as before, that at all search steps k the distributions $\mathbf{Q}_{s,a}^{\mathrm{tree},k}$ were expressed as beta distributions; that these distributions were, at each step, independent of one another for different states or actions, as were the posterior distributions over transition and reward functions; and that the distribution of the maximum (with respect to action choice) of a state-action value was the distribution corresponding to the single, apparently optimal action. See Mannor et al.[5] for analysis and experiments on the accuracy of a similar approach.

In particular, we initialized the 0-step value distribution $\mathbf{Q}_{s,a}^{\mathrm{tree},0}, \forall(s,a)$ as equal to the beta distribution over reward probability $R(s)$. We did this for both terminal and nonterminal states — the immediate reward distribution for nonterminal states was determined from the same prior by conditioning on the absence of reward each time the state is visited since in our simplified MDPs, reward is only available at terminal states. Then, for nonterminal states s, we repeatedly searched a further step down the tree, estimating the k-step value distributions $\mathbf{Q}_{s,a}^{\mathrm{tree},k}$ as a function of the $k{-}1$-step value distributions $\mathbf{Q}_{s,a}^{\mathrm{tree},k\text{–}1}$. As before, a $\mathbf{Q}$ distribution was approximated by the beta distribution matching the mean and variance of the (complicated) exact distribution. These are just the moments of Equation 1 (which describes the probability of future reward if the transition and successor state value functions were known) with respect to the distributions over those functions. After action a in state s, the probabilities $t_1 \dots t_n$ of transitioning to states $s_1 \dots s_n$ are Dirichlet-distributed, and the successor states' values $q_1 \dots q_n$ are beta-distributed (each state's as $\mathbf{Q}^{\text{tree},k\text{–}1}_{s_i,a_i}$ for the apparently best action a_i). Taking into account our independence assumptions, the mean and variance of $\mathbf{Q}^{\text{tree},k}_{s,a}$ are:

$$\mu^{\text{tree},k}_{s,a} = \sum_{i=1}^{n} \langle t_i \rangle \langle q_i \rangle \tag{5}$$

$$(\sigma^2)^{\text{tree},k}_{s,a} = \sum_{i,j:i \neq j} \langle t_i t_j \rangle \langle q_i \rangle \langle q_j \rangle + \sum_{i=1}^{n} \langle t_i^2 \rangle \langle q_i^2 \rangle - (\mu^2)^{\text{tree},k}_{s,a} \tag{6}$$

where the bracketed values are standard Dirichlet and beta moments for the distribution over T and the successor state $\mathbf{Q}$ distribution. This iteration was repeated until the distributions converged.

In more realistic domains with many states, it is impractical to re-compute value distributions at every state. This can be addressed by a number of methods, including pruning (examining only certain paths out of each state at each step). We modeled the "computational noise" or inaccuracy that would result as extra variance accumulating over each step of tree search. In particular, at each step k, we added a penalty to the variance $(\sigma^2)^{\text{tree},k}_{s,a}$ of a constant ν times the probability that the successor state s' was nonterminal.

To model reward devaluation, we replaced the distribution over $R(s)$, the reward probability for the terminal state corresponding to the devalued outcome, reducing its expected value. Through the value iteration, this immediately impacted all subsequently computed estimates $\mathbf{Q}^{\mathrm{tree}}_{s,a}$.

Noise model, priors, and parameters

The final complexity is that we assumed a nonstationary task — that is, that the MDP functions T and R could change randomly over time. Rather than employing an explicit, generative model of change, we captured nonstationarity using an exponential forgetting heuristic, whereby at each timestep, the parameters defining the cache system's distributions $\mathbf{Q}^{\mathrm{cache}}$ and the tree system's distributions over transition and reward functions decayed exponentially (with factor γ) toward their respective priors at each timestep. Such decay captures the decline in relevance of past samples given possible intervening change. As the decay factors were matched between controllers (as were the priors), this corresponds to equivalent time horizons on past data — i.e., equivalent assumptions about the speed of change of the MDP.

While the qualitative effects we demonstrated are robust, our theory has a number of free parameters. One advantage of a normative approach is that these are often not arbitrary quantities (like a "learning rate") but rather assertions about regularities in the external environment, such as how quickly tasks change. Thus, although they are at present chosen rather arbitrarily, they suggest directions for future experimental test.

Parameters for our simulations were as follows. The softmax parameter β was 5. The tree system's prior over the transition functions was a symmetric Dirichlet with parameter $\alpha = 1.0$ and over the reward functions was $Beta(0.1, 0.1)$ (encoding a prior assumption that outcome utilities are likely deterministic). The cache system's priors over the Q functions were matched: $Beta(0.1, 0.1)$ for terminal states, and for nonterminal states the same beta distribution implied by tree search on the tree system's prior over MDPs. The step penalty ν was 0.005, and the reward distribution for a devalued outcome in both tree and cache systems was $Beta(1, 15)$. The exponential forgetting factor γ was 0.98.

Simulations were run 250-1,000 times and means reported (results vary between runs due to stochastic action choice). In all cases, confidence intervals on the plotted quantities (s.e.m.) were too small to visualize; error bars were thus omitted.

References

1. Dearden, R., Friedman, N. & Russell, S.J. Bayesian Q-learning. in *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI)* 761–768 (1998).

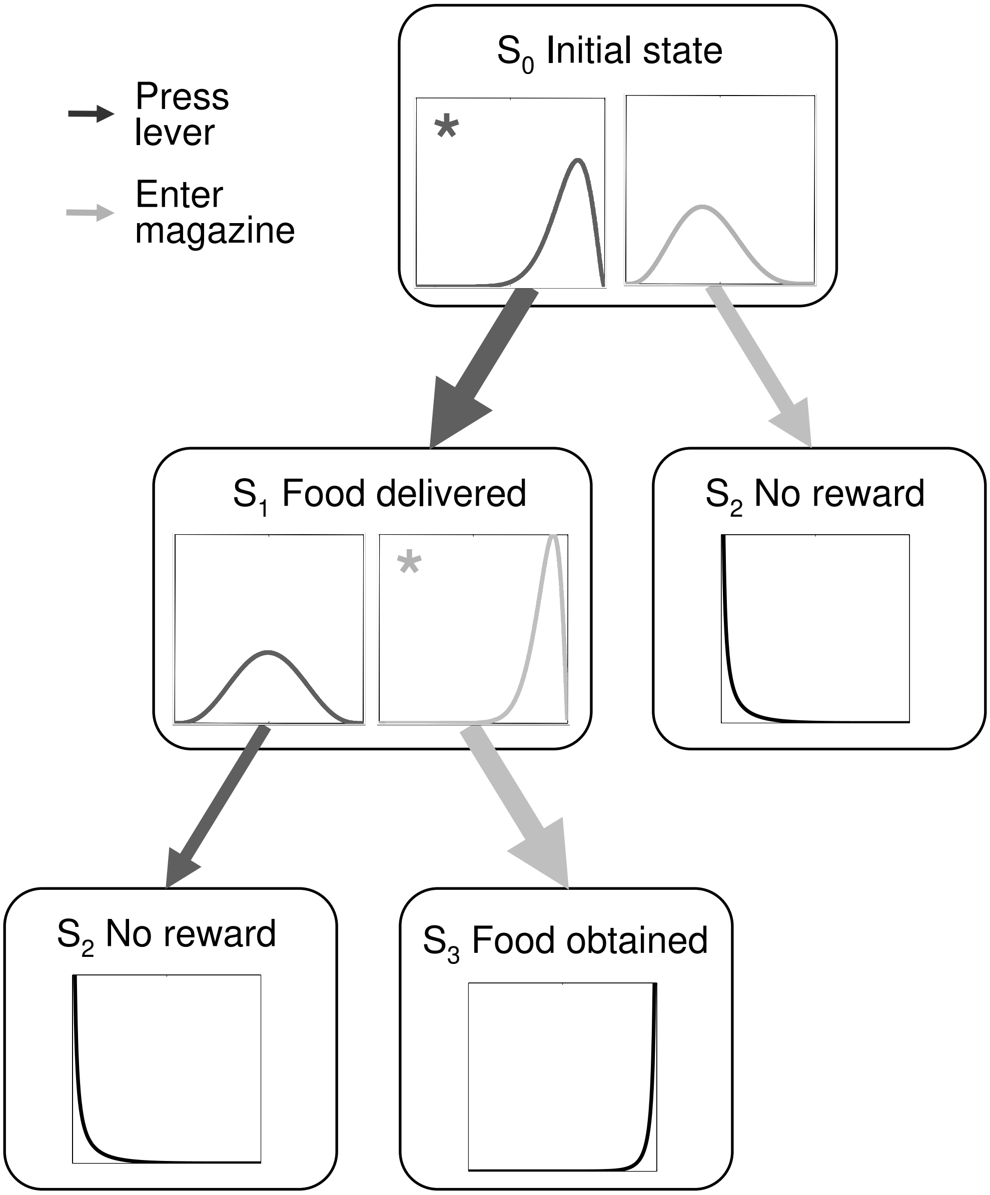
2. Engel, Y., Mannor, S. & Meir, R. Bayes meets Bellman: The Gaussian process approach to temporal difference learning. in *Proceedings of the 20th International Conference on Machine Learning (ICML)* 154–161 (2003).

3. Dearden, R., Friedman, N. & Andre, D. Model based Bayesian exploration. in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)* 150–159 (1999).

4. Strens, M. A Bayesian framework for reinforcement learning. in *Proceedings of the 17th International Conference on Machine Learning (ICML)* 943–950 (2000).

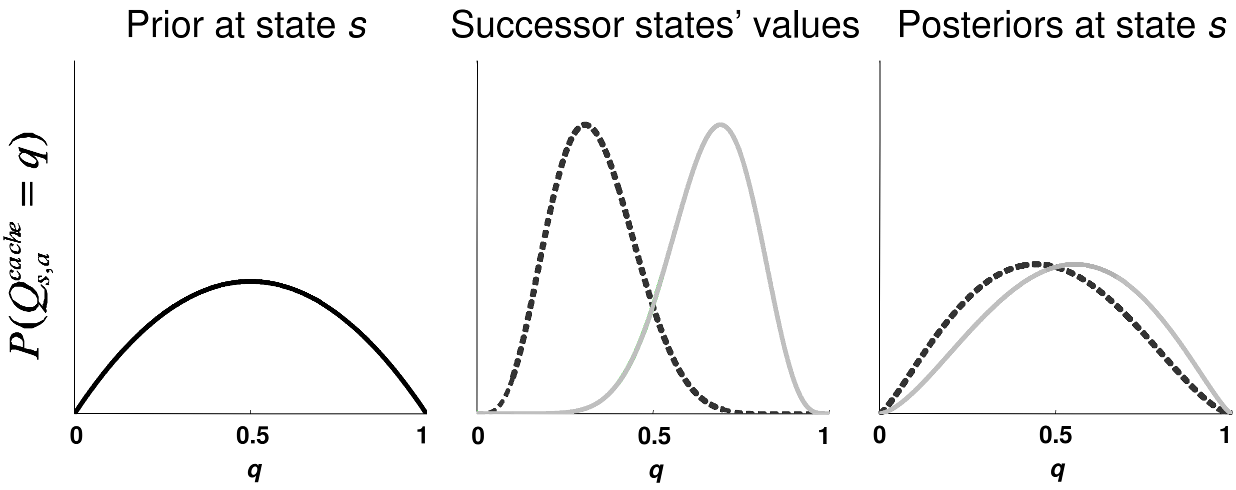
5. Mannor, S., Simester, D., Sun, P. & Tsitsiklis, J.N. Bias and variance in value function estimation. in *Proceedings of the 21st International Conference on Machine Learning (ICML)* 568–575 (2004).

6. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).



Supplementary Figure 1

Value propagation in tree search, after 50 steps of learning the task in **Figure 1a**. The inset plots show the distributions over state-action values, $\mathbf{Q}_{s,a}^{\text{tree}}$, computed by the tree system from the learned distributions over the values of the terminal states (shown in black) and over the transition structure of the task. The distributions are plotted as the probability assigned to each possible value q. Their moments are given by iteration on Equations 5 and 6 in **Supplementary Methods** — each value distribution is a function of the value distributions for the best action (marked with an asterisk) at each possible successor state. Arrows represent the most likely transition for each state and action, and their widths are proportional to the likelihoods (the full set of mean transition probabilities is illustrated in **Fig. 4**). The better actions were better explored and hence more certain (narrower value distributions); distributions at each state were similar to the distributions at the most likely successor state, and more so when transition to that state was more likely. As iterations progressed backwards, distributions got broader.



Supplementary Figure 2

Example of learning in the cache algorithm, following a single transition from state s to s' having taken action a. The leftmost panel shows the prior distribution $\mathbf{Q}_{s,a}^{\text{cache}}$. The middle panel plots the distribution over the value of the successor state s' (specifically, the distribution $\mathbf{Q}_{s',a'}^{\text{cache}}$ for the best action a' in s'). The curves illustrate two different successors, a more and less favorable state. For both of the successor states, the right panel plots the posterior distribution (whose moments are given by Equations 3 and 4 in **Supplementary Methods**) over the original state and action, $\mathbf{Q}_{s,a}^{\text{cache}}$, as updated following the visit to s'. The effect of learning was to nudge the predecessor state's value distribution in the direction of the value of the successor state.