CONNECTIONS BETWEEN COMPUTATIONAL AND NEUROBIOLOGICAL PERSPECTIVES ON DECISION MAKING

Decision theory, reinforcement learning, and the brain

PETER DAYAN
University College London, London, England

AND

NATHANIEL D. DAW
New York University, New York, New York

Decision making is a core competence for animals and humans acting and surviving in environments they only partially comprehend, gaining rewards and punishments for their troubles. Decision-theoretic concepts permeate experiments and computational models in ethology, psychology, and neuroscience. Here, we review a well-known, coherent Bayesian approach to decision making, showing how it unifies issues in Markovian decision problems, signal detection psychophysics, sequential sampling, and optimal exploration and discuss paradigmatic psychological and neural examples of each problem. We discuss computational issues concerning what subjects know about their task and how ambitious they are in seeking optimal solutions; we address algorithmic topics concerning model-based and model-free methods for making choices; and we highlight key aspects of the neural implementation of decision making.

The abilities of animals to make predictions about the affective nature of their environments and to exert control in order to maximize rewards and minimize threats to homeostasis are critical to their longevity. Decision theory is a formal framework that allows us to describe and pose quantitative questions about optimal and approximately optimal behavior in such environments (e.g., Bellman, 1957; Berger, 1985; Berry & Fristedt, 1985; Bertsekas, 2007; Bertsekas & Tsitsiklis, 1996; Gittins, 1989; Glimcher, 2004; Gold & Shadlen, 2002, 2007; Green & Swets, 1966; Körding, 2007; Mangel & Clark, 1989; McNamara & Houston, 1980; Montague, 2006; Puterman, 2005; Sutton & Barto, 1998; Wald, 1947; Yuille & Bülthoff, 1996) and is, therefore, a critical tool for modeling, understanding, and predicting psychological data and their neural underpinnings.

Figure 1 illustrates three paradigmatic tasks that have been used to probe this competence. Figure 1A shows a case of prediction learning (Seymour et al., 2004). Here, human volunteers are wired up to a device that delivers variable strength electric shocks. The delivery of the shocks is preceded by visual cues (Cue A through Cue D) in a sequence. Cue A occurs on 50% of the trials; it is followed by Cue B and then a larger shock 80% of the time or

by Cue D and then a smaller shock 20% of the time. The converse is true for Cue C. Subjects can, therefore, in general expect a large shock when they get Cue A, but this expectation can occasionally be reversed. How can they learn to predict their future shocks? An answer to this question is provided in the Markov Decision Problem section; as described there, these functions are thought to involve the striatum and various neuromodulators. Such predictions can be useful for guiding decisions that can have deferred consequences; formally, this situation can be characterized as a Markov decision problem (MDP) as studied in the fields of dynamic programming (Bellman, 1957) and reinforcement learning (Sutton & Barto, 1998).

Figure 1B depicts a decision task that is closely related to signal detection theory (Green & Swets, 1966) and has been particularly illuminating about the link between neural activity and perception (Britten, Newsome, Shadlen, Celebrini, & Movshon, 1996; Britten, Shadlen, Newsome, & Movshon, 1992; Gold & Shadlen, 2001, 2002, 2007; Shadlen, Britten, Newsome, & Movshon, 1996; Shadlen & Newsome, 1996). In the classical version of this task, monkeys watch a screen that shows moving dots. A proportion of the dots is moving in one direction; the rest are moving in

P. Dayan, dayan@gatsby.ucl.ac.uk
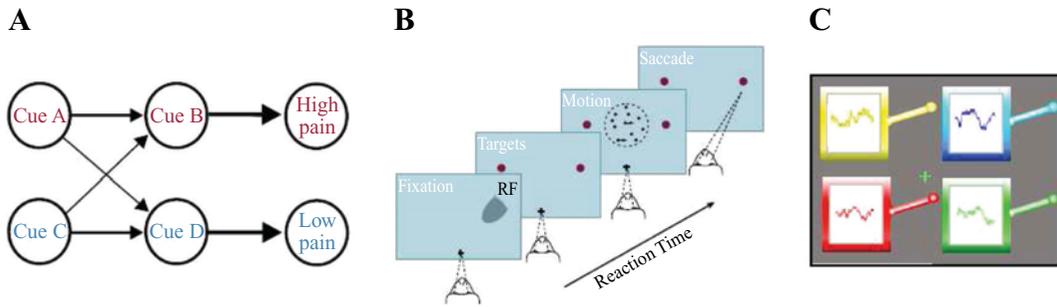
A   **B**   **C**

Figure 1. Paradigmatic tasks. (A) Subjects can predict the magnitude of future pain from partially informative visual cues that follow a Markov chain (Seymour et al., 2004; see the Markov Decision Problem section). (B) Monkeys have to report the direction of predominant motion in a random-dot kinematogram by making an eye movement (Britten, Shadlen, Newsome, & Movshon, 1992); see the Signal Detection Theory section. In some experiments, the monkeys have the additional choice of whether to act or collect more information (Gold & Shadlen, 2007); see the Temporal State Uncertainty section. (C) Subjects have to choose between four evolving, noisy bandit machines (whose payments are shown in the insets) and, so, must balance exploration and exploitation (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006); see the Exploration and Exploitation section.

random directions. The monkeys have to report the coherent direction by making a suitable eye movement. By varying the fraction of the dots that moves coherently (called the *coherence*), the task can be made easier or harder. The visual system of the monkey reports evidence about the direction of motion; how should the subject use this information to make a decision? In some versions of the task, the monkey can also choose *when* to emit its response; how can it decide whether to respond or to continue collecting information? These topics are addressed in the Signal Detection Theory and Temporal State Uncertainty sections, along with the roles of two visual cortical areas (MT and lateral intraparietal area [LIP]). The simpler version can be seen as a standard signal detection theory task; the more complex one has been analyzed by Gold and Shadlen (2001, 2007) as an optimal-stopping problem. This, in turn, is a form of partially observable MDP (POMDP) related to the sequential probability ratio test (SPRT; Ratcliff & Rouder, 1998; Shadlen, Hanks, Churchland, Kiani, & Yang, 2007; Smith & Ratcliff, 2004; Wald, 1947).

Finally, Figure 1C shows a further decision-theoretic wrinkle in the form of an experiment on the trade-off between exploration and exploitation (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). Here, human subjects have to choose between four one-armed bandit machines whose payoffs are changing over time (shown by the curves inside each). The subjects can find out about the current value of a machine only by choosing it and, so, have to balance picking the machine that is currently believed best against choosing a machine that has not recently been sampled, in case its value has increased. Problems of this sort are surprisingly computationally intractable (Berry & Fristedt, 1985; Gittins, 1989); the section of the present article on Exploration and Exploitation discusses the issues and approximate solutions, including one that, evidence suggests, implicates the fronto-polar cortex.

Despite the apparent differences between these tasks, they actually share some deep underlying commonalities. In this review, we provide a straightforward formal framework that shows the links, give a computationally minded view

of the method for solving the problems, and discuss these particular cases and their near relatives in some detail. A wealth of problems and solutions that has arisen in different areas of psychology and neurobiology is thereby integrated, and common solution mechanisms are identified. In particular, viewing these problems as different specializations of a common task involving both sensory inference and learning
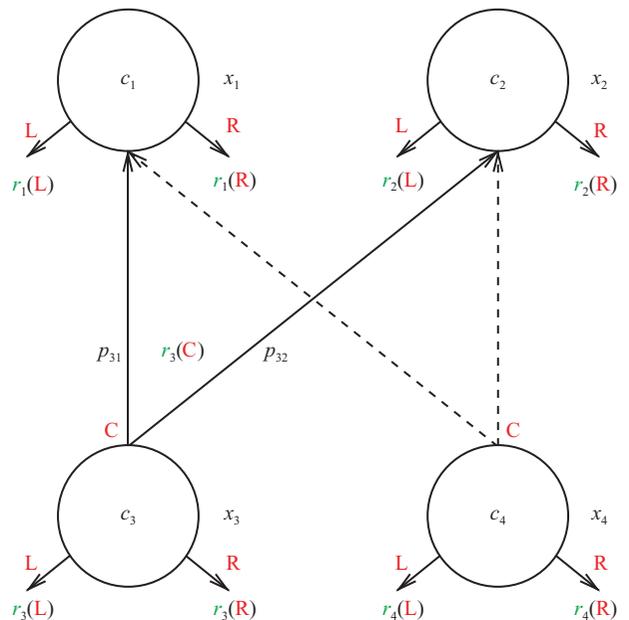

Figure 2. An abstract decision-theoretic task. Subjects normally start in state x_3 or x_4, signaled by cues c_3 or c_4. They have three options (L, R, or C): The former two lead to rewards or punishments such as $r_3(L)$; the latter lead via stochastic transitions (probabilities such as p_{31}) to states x_1 or x_2, which are themselves signaled by cues c_1 and c_2, and license rewarded or punished choices L and R. Subjects can be (partially) ignorant about their states if the cues are confusable (e.g., if x_3 "looks like" x_4), and/or about the rules of the task (the rewards and probabilities). In some cases, the subjects might start in x_1 or x_2. Different options generate a wide family of popular decision-theoretic problems.

components gives strong clues as to how sensory systems and computational mechanisms involved in the signal detection tasks—such as areas MT and LIP—are likely to interact with the basal ganglia and neuromodulatory systems that are implicated in the reinforcement learning tasks.

We tie the problems together by inventing a new, slightly more abstract assignment (shown in Figure 2). Particular specializations of this abstraction are then isomorphic to the tasks associated with Figure 1. The case of Figure 2 is an apparently simple maze-like choice task that we might present to animal or human subjects, who have to make decisions (here, choices between actions L, R, and C) in order to optimize their outcomes (r). Optimal choices involve balancing current and future rewards and costs and handling different forms of uncertainty about the rules of the task and the state within it.

Two critical dimensions that emerge from a consideration of Figure 2 concern what the subjects *know* and what they are trying to *accomplish*. The prediction task shown in Figure 1A arises when subjects are ignorant of the rules of the task but know their *state* or situation within it. Conversely, the psychophysical discrimination tasks shown in Figure 1B originate in a case in which subjects know the rules of the task but are only incompletely certain about the state. The exploration/exploitation trade-off shown in Figure 1C can be seen as combining both of these in a case in which subjects are ambitious about behaving optimally in the face of whatever uncertainty they have. Critically, through the medium of the task shown in Figure 2, all these problems can be characterized as requiring common computations. Realizing the computations leads to algorithmic issues having to do with different ways in which information from past and present trials can be accumulated and, thence, to implementational issues in terms of the neural structures involved in the solutions.

FOUNDATIONAL ISSUES

The task shown in Figure 2 involves only four choice points or *states* (x_1, x_2, x_3, and x_4) signaled, perhaps imperfectly (i.e., leaving some uncertainty, in a way we will formulate precisely later), by cues (c_1, c_2, c_3, and c_4). Three actions are possible (L, R, and C) at the states, and it is the choices between these that the subjects must make. The choices lead to rewards or punishments (with values or utilities r, which depend on the states and actions), and/or to transitions from one state to the next (x_3 to x_1 or x_2, etc.). We consider that single trials end when an actual outcome is achieved; the subjects then start again. In general, subjects' choices may be only probabilistically related to the outcomes. In the standard case for this review, we will have the rewards being biggest for L at x_3 and x_1 and for R at x_4 and x_2, with C being costly. The subjects may not know these payoffs at the outset.

Computational Issues
As has been mentioned, there are two main dimensions defining the problem for the subjects: one having to do with what they are assumed to *know* about the task, the other defining the nature of their *ambition*.

In terms of knowledge, the subjects might be ignorant of their precise state in the problem (i.e., which x_i they currently occupy), and/or the rules of the task (i.e., the transition probabilities and rewards contingent on particular actions in states). If they know both in the standard version of the task, which requires L at x_3 and x_1 and R at x_4 and x_2, with C being costly, and they know that they are at x_3, say, they should choose L. However, if they know the rules and know that they are at either x_3 or x_4, but do not know which for sure (perhaps because the cues c_3 and c_4 are similar or identical), it might be worth the cost of choosing C in order to collect information from c_1 and c_2 (if these are more distinct), the better to work out which action is then best.

The problems of balancing such costs and benefits get much harder if the subjects might not even completely know the rules of the task. This is necessarily the case at the outset of animal experiments and is also more persistently true when, as is common in experiments, the rules are changed over time. In these cases, subjects will have to learn the rules from experience. However, experience will normally only partly specify the rules, leaving some ignorance and uncertainty, and it will often be important to take proper account of this.

Second, in terms of their ambition, the subjects might have the modest goal of exploitation—that is, trying to make the reward for the current trial as valuable as possible, given whatever they currently know about the task. In the case in which the subjects start at x_3 or x_4, this involves comparing rewards available for the immediate choice of L or R with the integrated cost of C and the subsequent reward from L or R at x_1 or x_2. How to trade off immediate and deferred reward optimally depends on subjects' preferences with respect to temporal discounting (e.g., Ainslie, 2001; Kable & Glimcher, 2007; McClure, Laibson, Loewenstein, & Cohen, 2004).

More ambitious subjects might seek to combine exploration and exploitation. That is, they might look to make every single choice correctly in the light of the fact that not only might it lead to a good outcome on this trial, but also it could provide information that will lead the subject to be more proficient at getting better outcomes in the future. This goal—choosing so as to maximize the integrated rewards obtained over many trials throughout the course of learning, trading off the immediate benefits of exploitation and the deferred benefits of exploration—is sometimes called *lifetime optimality*. Again, how these are balanced depends on temporal discounting.

Note that these two computational dimensions are not wholly independent; for instance, given complete knowledge of rules and state, exploration is moot.

Algorithmic Issues
Different points along the combined computational dimensions lead to a wide variety of different problems. Some of these are formally tractable—that is, have algorithms that require only moderate amounts of memory space or time to compute optimal solutions. Other points, particularly those involving incomplete knowledge or lifetime optimality, are much more challenging and typically require approximations, even for nonneurobiological systems.

Algorithms differ in how they draw on experience to estimate quantities relevant to the decision and in how they render these into choices. The most important algorithmic dimension is that distinguishing *model-based* and *model-free* methods (Sutton & Barto, 1998). Crudely speaking, model-based methods make explicit use of the actual, or learned, rules of the task to make choices. Importantly, even when the rules are fully known, it takes some computation to derive the optimal decision for a particular state from these more basic quantities.

Model-free methods eschew the rules of the task and, instead, use and/or learn putatively simpler quantities that are sufficient to permit optimal choices. For instance, in Figure 2, given complete knowledge of the state, it clearly is enough just to know four letters—namely, the best choices at $x_1 \ldots x_4$. This is an example of a policy. Obtaining it in the face of ignorance of the rules lies at the heart of reinforcement learning methods. Policies can be learned directly or can be derived from other information, such as the expected future utilities (*values*) that will accrue from different actions or states. Policies can also be derived (in model-based methods) from the rules of the task. Indeed, one of the most important products of the field of reinforcement learning (Sutton & Barto, 1998) is a range of model-free algorithms for solving the exploitation problem.

When the observations or cues do not precisely pin down the state, a policy mapping states to actions is obviously of little use. Given a model of the rules, including those relating states such as x_3 to cues such as c_3, the beliefs about the current state (called the *belief state*) can be calculated on the basis of the observations. The belief state can then, in a formal sense, stand in for the true state, so the policy becomes a function of this instead. This substitution of belief state for state is a recurring theme in the solution to the tasks discussed below.

Implementational Issues

It has long been suggested that there is a rather direct mapping of model-free reinforcement learning algorithms onto the brain, with the neuromodulator dopamine serving as a teaching signal to train values or policies by controlling synaptic plasticity at targets such as the ventral and dorsolateral striatum (Barto, 1995; Daw, Niv, & Dayan, 2005; Friston, Tononi, Reeke, Sporns, & Edelman, 1994; Joel, Niv, & Ruppin, 2002; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Suri & Schultz, 1998; Wickens, 1990). For aversive outcomes such as the shocks shown in Figure 1A, there is much less evidence about the overall neural substrate. More recently, it has been suggested that the brain also employs model-based methods for planning under uncertainty about the rules, in a different set of circuits involving the prefrontal cortex and the dorsomedial striatum (Balleine, Delgado, & Hikosaka, 2007; Daw et al., 2005; Dickinson & Balleine, 2002; Everitt & Robbins, 2005).

Most of this work has focused on rule, value, or policy learning, ignoring the issue of state uncertainty; indeed, arguably, the primary obstacle toward employing either the model-based or the model-free methods in a real-world context is the gulf between the highly constrained and refined state abstraction on which these theories rely and the rich, multifarious, and ambiguous sensory world actually facing an organism.

Conversely, model-based methods for state estimation from noisy sensory input have been extensively investigated in a rather different set of psychophysical tasks (Britten et al., 1996; Britten et al., 1992; Gold & Shadlen, 2007; Parker & Newsome, 1998; Platt & Glimcher, 1999), exemplified by Figure 1B, focusing on the mapping of input information coded in sensory regions into decision-theoretic quantities coded in more motor-associated regions of the cortex. However, this work has generally confined itself to fairly rudimentary and limited forms of learning.

The aim of the remainder of this article is to situate both learning and state estimation mechanisms in a single framework.

We will use the examples shown in Figure 1, rendered in the abstract forms shown in Figure 2, to illustrate the key general principles established above. The examples cover neural reinforcement learning (Montague et al., 1996; Schultz et al., 1997), Bayesian psychophysics (Britten et al., 1992; Shadlen & Newsome, 1996), information gathering and optimal stopping (Gold & Shadlen, 2001, 2007), and the exploration/exploitation trade-off (Daw, O'Doherty, et al., 2006). The first example develops basic reinforcement learning methods for tasks in which the state is known; the rest exemplify how these can be extended, with belief states as the formal replacement for uncertain true states.

In each section, we first will describe the formal computational notions and ideas, then the relevant algorithmic methods associated with the computations, and finally the psychological and neurobiological tasks and mechanisms that are implicated.

THE MARKOV DECISION PROBLEM

The central problem in the prediction task shown in Figure 1A is that until the subject observes Cue A or C at the start of a trial, he or she does not know whether he or she will receive a shock at the end of the trial; the cue makes the outcome more predictable. In Markov problems—that is, domains in which only the current state matters, and not the previous history—there turns out to be a computationally precise way of defining the goal for predicting future reinforcers. When a choice of actions is available, the goal also provides a formalization of optimal selection. There is also a variety of algorithmic methods for acquiring predictions and using the predictions for control.

The Computational Problem

Consider the case shown in Figure 3A, in which the cues unambiguously identify the state (c_1 for x_1, and so on). We will first consider decision making when the rules are given and then move onto the standard reinforcement learning problem in which the rules of the task are unknown and the subject must discover how best to behave by trial and error.

Given the rules, the task for the subject is simply to work out the best *policy*, π_i^* (the asterisk identifying it as being best), which specifies an assignment of an action
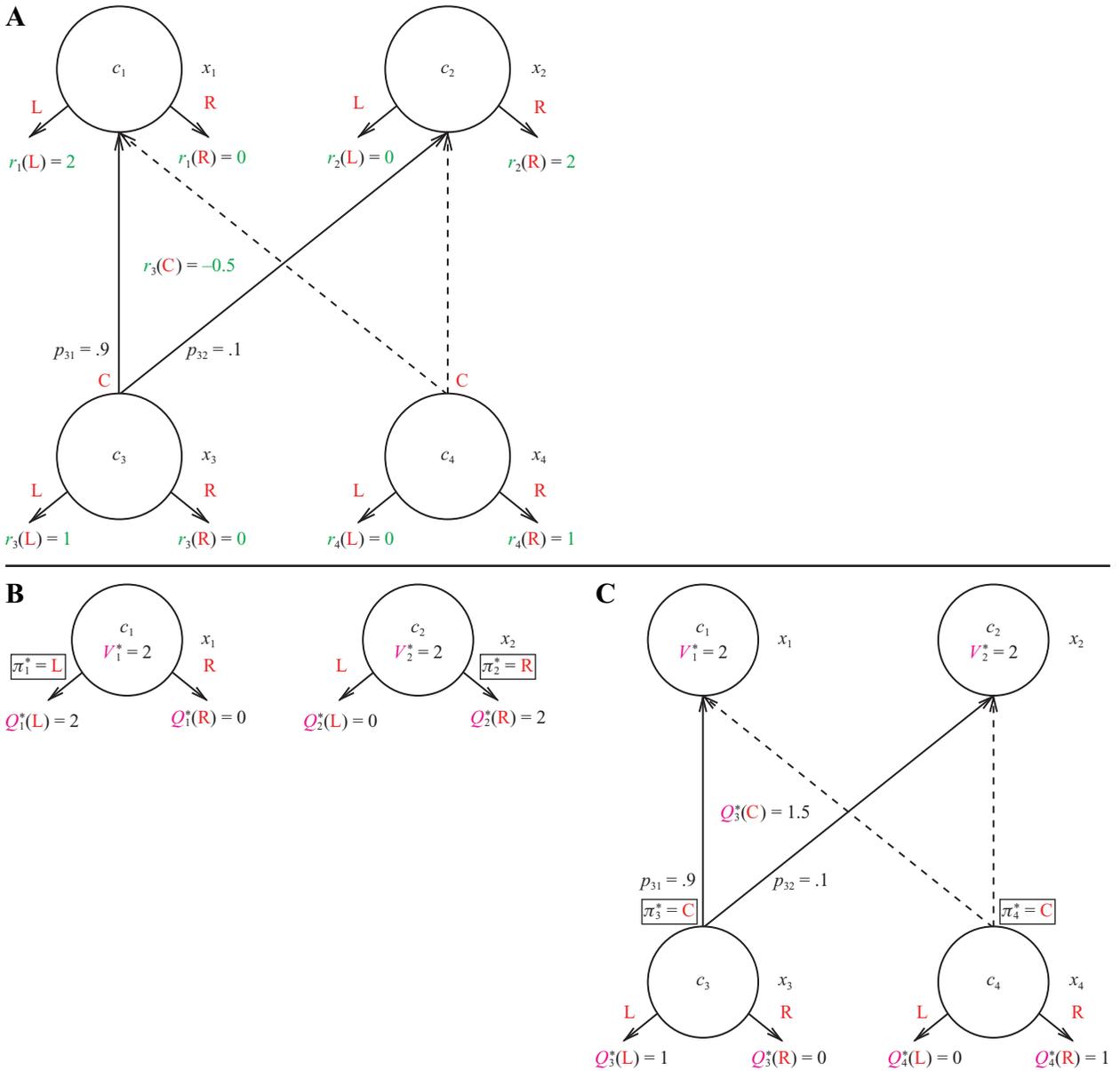
A

B

C

Figure 3. Markov decision problem (MDP). (A) The version of the basic task rendered as a simple MDP. Each state x_i has a distinct cue c_i, and the rewards and transition probabilities are as shown (the case for C at x_4 is symmetric with respect to x_3). The solution to MDP involves optimal state-action values $Q_i^*(a)$, state values V_i^*, and thence the optimal policy π_i^* (shown in the boxes) first for states x_1 and x_2 (B), and then for x_3 and x_4 (C). The calculation for x_3 and x_4 depends only on V_1^* and V_2^* and not on the manner by which the reward from x_1 or x_2 is achieved.

$a \in \{L, R, C\}$ to each state x_i. The probabilities $p_{ij} = p_{ij}(C)$ indicate the probabilities of going from state x_i to x_j under action C, whose cost is $r_i(C)$; actions L and R have deterministic consequences. Exactly how the "best" policy is defined depends on the particular goal. For now, we will assume that the rewards and costs earned across each whole single trial are simply summed, and it is this sum that has to be predicted and optimized.

First, consider the case in which the goal is exploitation within a single trial, to maximize the average, or expected, reward. If we consider state x_1, the task is straightforward;

the value of each action $Q_1^*(a)$, defined as the expected return for performing that action, is

$$Q_1^*(a) = r_1(a), \tag{1}$$

and a best action (i.e., one that maximizes this expected return) is

$$\pi_1^* = \text{argmax}_{a \in \{L,R\}} [Q_1^*(a)] = L, \tag{2A}$$

and similarly

$$\pi_2^* = \text{argmax}_{a \in \{L,R\}} [Q_2^*(a)] = R. \tag{2B}$$

All these quantities are shown in Figure 3B (the best actions are boxed).

The case for x_3 and x_4 is more complicated, since if C is chosen, it would seem that one should consider not only the action there, but also the subsequent action at x_1 or x_2, since the rewards associated with each action are to be summed. Critically, the task has a Markov structure, meaning that *how* the subject gets to x_1, for instance, does not bear at all on the choice to be made at that point and the rewards that will subsequently accrue (the Markovian mantra is that "the future is independent of the past, given the present state"). The theory of dynamic programming (Bellman, 1957; Bertsekas, 2007; Ross, 1983) fashions this observation into computational and algorithmic methods, which themselves underlie reinforcement learning. The idea is that states x_1 and x_2 also get values V_i^* under the optimal policies there, defined as the best possible expected return starting there

$$V_1^* = \max_{a \in \{L, R\}} [Q_1^*(a)] = 2 \qquad (3A)$$

and similarly at x_2

$$V_2^* = \max_{a \in \{L, R\}} [Q_2^*(a)] = 2, \qquad (3B)$$

which are shown in Figure 3B. These values will be available, provided the subjects choose correctly (i.e., according to π_i^* if they get to those states), and thus can act as surrogate rewards for reaching the respective states, hiding all the complexity of how those rewards are achieved. Then, we can write the value of choosing C at x_3 as

$$Q_3^*(C) = r_3(C) + \sum_j p_{3j} V_j^* = r_3(C) + p_{31} V_1^* + p_{32} V_2^*, \quad (4)$$

since the reward for the first action (at state x_3; $r_3(C)$) is added to that for the second action (at state x_1 or x_2). The probabilities p_{31} and p_{32} arise because the value of the action is the expected value of doing the action. These quantities determine the probabilities of the transitions to States 1 and 2, and multiply the values V_1^* and V_2^*, which indicate the rewards that can be achieved starting from those states, given appropriate actions there.

The values of L and R at x_3 are just $Q_3^*(L) = r_3(L)$ and $Q_3^*(R) = r_3(R)$, and so the optimal policy at x_3 is, as in Equations 2A and 2B,

$$\pi_3^* = \mathrm{argmax}_{a \in \{L, R, C\}} [Q_3^*(a)] = C \qquad (5A)$$

and similarly

$$\pi_4^* = \mathrm{argmax}_{a \in \{L, R, C\}} [Q_4^*(a)] = C. \qquad (5B)$$

That is, in the example in Figures 3A–3C, it is optimal to take action C at x_3 (and x_4), since $1.5 = r_3(C) + r_1(L) > r_3(L) = 1$. That this is true, of course, depends on the precise values of the rewards.

Although we showed only the computations underlying the simplest dynamic programming problem, solving more realistic cases that, nevertheless, retain the structure of our task is a straightforward extension (see, e.g., Bertsekas, 2007; Sutton & Barto, 1998). The central requirements are that the states and rules are known, that rewards are additive over time (although future rewards can be discounted exponentially, as if by an interest rate), and that the problem is Markovian, so future transitions and rewards depend only on the current state, and not on the path the subject took to achieve the current state. The constraint that the state satisfy this Markov property is critical to all the analysis above and, concomitantly, is a major hurdle in connecting this abstract formalism with more realistic situations in which the cues do not determine the state. This issue is central to the remaining examples in the article.

We have so far assumed that the subjects know the rules of the task. In the Exploration and Exploitation section, we will consider the (much more involved) computational problem that arises in the case in which the subjects do not know the rules and have the ambition of optimally balancing exploration to find them out and exploitation of what they already know. However, there is an important intermediate case in which the subjects do not know the rules and, so, have to learn them from experience, but in which their only ambition is exploitation—that is, doing as well as possible in the current trial, ignoring the future. The standard way to conceive of experience here is in terms of sampling states, transitions, and rewards—for instance, starting at a state (x_3), choosing an action (C), experiencing a transition (probabilistically to x_1 or x_2), and receiving a reward ($r_3(C)$, which in general could also be stochastic). Learning about the task from such samples is, of course, an example of a very general statistical estimation problem on which we can only touch. The critical consequence that we will consider, though, is that given only limited numbers of samples, subjects will be making choices in the face of uncertainty about the task. Even without considering exploration to reduce the uncertainty, exploitation can be significantly affected by it.

Algorithmic Approaches

The theory of dynamic programming (Bellman, 1957; Ross, 1983) specifies various algorithms for calculating the optimal policy—notably, *policy* and *value* iteration (Bertsekas, 2007; Puterman, 2005; Sutton & Barto, 1998). The key observation is that Equation 4, which is one of a number of forms of the so-called Bellman equation, specifies a consistency condition between the optimal Q^* values at one state (x_3) and those at other states (x_1 and x_2), via relationships such as Equations 3A and 3B. The different algorithms find optimal values by attacking any *inconsistencies*, but do so in different ways.

Standard dynamic programming algorithms are *model based*, in the sense that they solve the equations by making explicit use of the quantities $r_i(a)$ and $p_{ij}(C)$, following just the sort of reasoning described above. However, in the case of learning from experience in an unknown task, it is first necessary to acquire estimates of these from the samples. This can be relatively straightforward. For instance, consider the case in which the subject performs action C at state x_3 a total of M times, getting to states $x_{j^1} \ldots x_{j^M}$, for $j^1, \ldots, j^M \in \{1,2\}$ sampled from p_{31}; p_{32}, and experiencing rewards $r^1 \ldots r^M$, sampled from

$r_3(C)$. Then, one might estimate $r_3(C)$ and p_{31} using the sample mean estimates

$$\hat{r}_3(C) = \frac{1}{M} \sum_{k=1}^{M} r^k \qquad (6A)$$

and

$$\hat{p}_{31} = \frac{1}{M} \sum_{k=1}^{M} \chi(x_{jk}, x_1), \qquad (6B)$$

where χ is the characteristic function, meaning that the second sum just counts the number of times the transition was to x_1. Note that since the transitions are random (and the rewards, in general, might also be), these estimates involve sampling error. Bayesian estimates of the rules would quantify this error as uncertainty, a remark to which we will return when we consider exploration. For the purpose of exploitation, it is conventional simply to compute the optimal policy under the so-called *certainty-equivalent* assumption that the current estimates are correct.

By contrast with the model-based methods, whose calculations depend explicitly on the rules, there are various reinforcement learning algorithms that are *model free* and estimate values or policies directly using individual samples of rewards and state transitions in place of estimated average rewards or state transition probabilities. One family of methods, called *temporal difference algorithms* (Sutton, 1988), estimates values by computing a *prediction error* signal measuring the extent to which successively predicted values and sampled rewards fail to satisfy the consistency prescribed by Equation 4. A typical prediction error, derived from the difference between right- and left-hand sides of Equation 4, is

$$\delta^k = r_3(C) + V_{jk}^* - Q_3^*(C) \qquad (7)$$

for trial k, in which the transition went from state x_3 to state x_{jk}. Such errors can be used to update value estimates (in this case of $Q_3^*(C)$) to reduce inconsistencies. These algorithms are guaranteed to converge to optimal value functions (which determine optimal policies) in the limit of indefinite sampling (Jaakkola, Jordan, & Singh, 1994; Bertsekas & Tsitsiklis, 1996; Watkins, 1989). However, given only finite experience, the value estimates will again be subject to sampling noise, and decisions derived from them may, therefore, be incorrect. Model-free methods of this sort are sometimes called *bootstrapping* methods, since they change estimates (here, of $Q_3^*(C)$) on the basis of other estimates (V_{jk}^*). This makes them statistically inefficient, since early on, estimates such as V_{jk}^* are inaccurate themselves and, so, can support only poor learning.

Although these algorithms are model free (i.e., not making explicit use of terms such as p_{31}), they are value based, since they work by estimating quantities such as $Q_i^*(a)$ and V_i^*, which are values of states and actions or of states, in terms of the summed reward that is expected to accrue across the whole rest of the trial. Apart from these value-based reinforcement learning algorithms, there is also a range of model-free methods that learn policies directly, without using the values as intermediate quantities. These policy-based methods (Baxter & Bartlett, 2001; Williams, 1992)

use stochastic policies (to allow sampling of the various options), which are parameterized, and adjust the parameters using learning rules that perform a form of stochastic gradient ascent or hill-climbing on the overall expected reward. For instance, consider the case of state x_1. We can represent a stochastic parameterized policy there as

$$P_1(a = L; w_1) = \pi_1(w_1) = \sigma(w_1), \qquad (8)$$

where

$$\sigma(w) = \frac{1}{[1 + \exp(-w)]}$$

is the standard logistic sigmoid function. Naturally, $P_1(a = R; w_1) = 1 - \pi_1(w_1) = \sigma(-w_1)$. Here, the greater w_1, the more likely the subject is to choose L at state x_1. The average reward based on this policy is

$$\langle r_1 \rangle_{w_1} = \sigma(w_1) r_1(L) + \sigma(-w_1) r_1(R).$$

If the subject employs this policy, choosing action a^k and getting reward r^k on trial k, and w_1 is *changed* according to a particular Hebbian correlation between the reward r^k and the probability of choice,

$$w_1 \Rightarrow w_1 + \varepsilon \Delta w_1^k, \qquad (9)$$

where

$$\Delta w_1^k = r^k \left[1 - P_1(a = a^k; w_1) \right],$$

then it can be shown that the average change in w_1 is proportional to the gradient of the expected reward:

$$\left\langle \Delta w_1^k \right\rangle_k \propto \frac{\partial \langle r_1 \rangle_{w_1}}{\partial w_1};$$

and so the latter quantity should increase on the average, at least provided that the learning rate ε, which governs how fast w_1 changes, is sufficiently small. Indeed, tight theorems delineate circumstances under which this rule leads in the end to the optimal policy π_1^*.

Rules of this sort are an outgrowth of some of the earliest ideas in animal behavioral learning—crudely suggesting that actions that are followed by rewards should be favored. As advertised, they work directly in terms of policies, not employing any sort of value as an intermediate quantity. They just require adapting so that they can, for instance, increase the probability of doing action C at state x_1 on the basis of the reward $r_3(L)$ that is available only at a later point.

Psychology and Neuroscience

There is a wealth of work in both psychology and neuroscience on tasks that can be considered to be Markov decision problems, and the exploitation problem has been a key focus. Importantly, the theory views predicting summed reinforcement as a key subproblem for decision making, and so much work has been concerned with predicting reinforcement in tasks without decisions.

Perhaps the best developed connection between these ideas and neural data is through prediction errors for model-free reinforcement learning, as in Equation 7. Ver-

sions of these have long played an important role in theories of behavioral conditioning—most famously, that of Rescorla and Wagner (1972). More recently, neural correlates of such error signals have been detected in a number of tasks and species.

Consider the experiment shown in Figure 1A. The task was designed to induce higher order prediction errors—that is, those arising from changes in expectations about future reinforcement, rather than from the immediate receipt (or nonreceipt) of a primary reinforcer. Such errors are characteristic of the bootstrapping strategy of temporal-difference algorithms, which take the changes in expectations (e.g., the difference between V_{jk}^* and Q_3^* in Equation 7) as signs of inconsistencies or errors in value predictions. Figure 4A highlights brain regions, notably in the ventral putamen (lateral striatum), where the measured BOLD signal was found to correlate over the task with a prediction error time series generated from a temporal-difference model.

For instance, Figure 4B shows the average BOLD signal from the right putamen on trials in which the subjects see Cue C followed by Cue B. In this case, the first cue indicates that a large shock is unlikely, but the later cue signals that it is certain. The change in expectation occasioned by the second cue induces a prediction error, reflected in increased BOLD activity. Conversely, Cue D following Cue A signals that a large shock previously thought to be likely will not occur; this is a negative prediction error (and a relative decrease in BOLD; see Figure 4C). Figure 4D illustrates how we can extend the same logic a further step back, just as in the dynamic programming analysis of MDPs. Here, since Cue A indicates that Cue B (and thence the large shock) is likely, it also induces a positive prediction error when it appears, signaling an end to the relatively safe period between trials.

Seymour et al.'s (2004) study was in the aversive domain. For appetitive outcomes, there is ample evidence that the phasic activity of dopamine cells in the ventral tegmental area and substantia nigra pars compacta in monkeys (e.g., Schultz, 2002), and the release of dopamine at striatal targets in rats (Day, Roitman, Wightman, & Carelli, 2007)

report quantities akin to the temporal difference prediction error in Equation 7. This comes on top of a huge body of results on the involvement of dopamine and its striatal projection in appetitive learning and appetitively motivated choice behavior (for some recent highlights, see Costa, 2007; Hyman, Malenka, & Nestler, 2006; Joel et al., 2002; Wickens, Horvitz, Costa, & Killcross, 2007). The proposal that this operates according to the rules of reinforcement learning (Balleine et al., 2007; Barto, 1995; Daw & Doya, 2006; Haruno et al., 2004; Joel et al., 2002; Montague et al., 1996; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; O'Doherty et al., 2004; Schultz et al., 1997; Suri & Schultz, 1998), in a way that ties together the at least equally extensive data on the psychology of instrumental choice with these neural data, has extensive, although not universal, support (e.g., Berridge, 2007; Redgrave, Gurney, & Reynolds, 2008).

However, behaviorally sophisticated experiments (reviewed in Balleine et al., 2007; Dickinson & Balleine, 2002) show that this is nothing like the whole story. These experiments study the effects of changing the desirability of rewards just before animals are allowed to exploit their learning. Model-based methods of control can use their explicit representation of the rules to modify their choices immediately in the light of such changes, whereas model-free methods, whose values change only through prediction errors (such as Equation 7), require further experience to do so (Daw et al., 2005). There is evidence for both sorts of control, with model-based choices (called *goal-directed* actions) dominating for abbreviated experience, certain sorts of complex tasks, and actions close to final outcomes. Model-free choices (called *habits*) are evident after more extensive experience, in simple tasks, and for actions further from outcomes. Furthermore, these two forms of control can be differentially suppressed by selective lesions of parts of the medial prefrontal cortex in rats (Killcross & Coutureau, 2003). Daw et al. (2005) argued that the trade-off between goal-directed actions and habits is computationally grounded in the differential uncertainties of model-based and model-free control in the light of limited sampling experience.
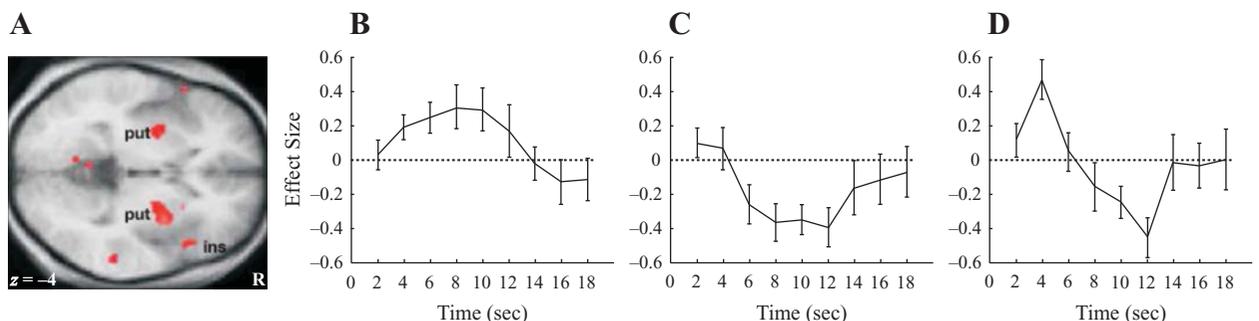

Figure 4. BOLD signals correlating with higher order prediction error in the aversive conditioning task shown in Figure 1A. (A) Regions of the bilateral ventral putamen (put; also the right anterior insula: ins) where the BOLD signal significantly correlated with prediction error. (B–D) BOLD time courses from the right putamen. (B) Positive prediction error: Cue B (contrasted against Cue D) following Cue C. (C) Negative prediction error: Cue D (contrasted against Cue B) following Cue A. (D) Biphasic prediction error: Cue A followed by Cue D, contrasted against Cue C followed by Cue B. From "Temporal Difference Models Describe Higher-Order Learning in Humans" by B. Seymour et al., 2004, *Nature, 429*, pp. 665, 666. Copyright 2004 by Nature Publishing Group. Adapted with permission.

Although more comprehensive, even this synthesis has an extremely limited scope. As was hinted above, the question of most relevance to the present review is how internally to create or infer a state space from just a booming, baffling confusion of poorly segmentable cues. That is, how to extract the equivalent of $x_1 \ldots x_4$, the underlying governors of the transitions and rewards, automatically from experience in an environment. The simplest versions of this issue are related to topics much more heavily studied in sensory neuroscience and psychophysics, and we now will turn to these.

SIGNAL DETECTION THEORY

The task shown in Figure 1B, in the version in which the subject cannot influence the length of time that the dots are shown, is one of sensory discrimination. Here, noisy and, therefore, unreliable evidence provided by motion-processing areas in the visual system has to be used to make as good a decision as possible to maximize reward. It maps onto the basic task in the case in which the rules are known, but the c_i inputs associated with the states are only partially informative about the states (because of the effects of noise).

Variants of this task—notably, ones involving the detection of a very weak sensory signal in the face of noise in the processing of input—are among the most intensively studied quantitative psychophysical tasks; it was because of this that they came to be used to elucidate the neural underpinnings of decision making.

The Computational Problem

Figure 5A shows the variant of the basic abstract problem that is a form of a classic signal discrimination task (Green & Swets, 1966). Here, the subject always starts in state x_1 or x_2, and the rules are assumed to be known, so that it is optimal to execute L in x_1 and R in x_2. However, the cues c_1 and c_2 are partly confusable; in other words, the subject observes c_α, which does not completely distinguish x_1 and x_2, and so it is uncertain which of the two states it occupies. This problem is sometimes called *partially observable*, or involving a *hidden state*, since the subject occupies a true state in the world that we write as $x_\alpha \in \{x_1, x_2\}$, but its

identity is at least partially hidden from the subjects. The subjective problem is illustrated in Figure 5B.

To formalize this task, it is necessary to specify the coupling between cues and states. The natural model of this involves the conditional distributions over the possible observations (the cues) given the states

$$p_1\left(c_\alpha = c \mid x_\alpha = x_1\right) \equiv p_1\left(c_\alpha \mid x_1\right)$$

and

$$p_2\left(c_\alpha = c \mid x_\alpha = x_2\right) \equiv p_2\left(c_\alpha \mid x_2\right),$$

which, in signal detection theory, are often assumed to be Gaussian, with means μ_1 and μ_2 (say, with $\mu_1 > \mu_2$) and variances σ_1^2 and σ_2^2. These distributions are shown in miniature in Figures 5A and 5B.

If the subject observes a particular c_α, then, given these distributions, what should it do? It needs a decision rule—a mapping, sometimes called a *test*—from its observation c_α to a choice of action, L or R.

There are four possibilities for executing one of these actions at one of the two states. Standard signal detection theory privileges one of the actions (say, L) and, thus, one of the states (here, x_1) and defines the four possibilities shown inside Table 1. Note that we could just as well have privileged R at x_2. Signal detection theory stresses the trade-off between pairs of these outcomes. For instance, subjects could promiscuously choose L despite evidence from c_α that x_2 is more likely. This would reduce misses, at the expense of introducing more false alarms.

Under Bayesian decision theory, subjects should maximize their expected reward, given the information they have received. The first step is to use the observation c_α to calculate the subjective *belief state*—that is, the posterior distribution over being in x_1 or x_2, given the data

$$P\left(x_\alpha = x_1 \mid c_\alpha\right) \equiv P\left(x_1 \mid c_\alpha\right) = \frac{p_1\left(c_\alpha \mid x_1\right)P\left(x_1\right)}{p\left(c_\alpha\right)}$$

$$= \frac{1}{1 + \dfrac{1}{l\left(c_\alpha\right)}\dfrac{P\left(x_2\right)}{P\left(x_1\right)}}, \quad (10)$$
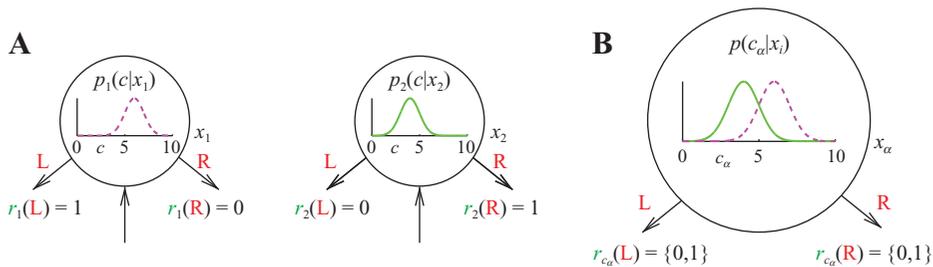

Figure 5. Signal detection theory task. Subjects start in $x_\alpha \in \{x_1, x_2\}$, but the cue c_α is confusing between these two possibilities, according to the distributions shown inside the states. **(A)** The objective state in the environment shows the consequence of choosing L or R at either of the two states. **(B)** The subjective state of the subject shows the confusion between the two possibilities; in particular (overloading the notation), the reward $r_{c_\alpha}(L)$ might be either $r_2(L) = 0$ or $r_1(L) = 1$. The distributions show $p_i(c_\alpha \mid x_i)$ (for the simple, equal-variance Gaussian case). Here, x_1, which requires L, is associated with slightly higher values of c_α than is x_2, which requires R.

Table 1
Signal Detection Contingencies

	Action	
Actual State	L	R
x_1	hit	miss
x_2	false alarm	correct rejection

where

$$l(c_\alpha) = \frac{p_1(c_\alpha \mid x_1)}{p_2(c_\alpha \mid x_2)}, \qquad (11)$$

is the so-called *likelihood ratio*, which indicates the relative chance that the observation c_α would have originated from x_1, versus x_2, and $P(x_1) = 1 - P(x_2)$ is the prior probability of starting in x_1. Figure 6A shows the logarithm of the likelihood ratio in the case in which the Gaussians have the same variance $\sigma_1 = \sigma_2$, and Figure 6B shows the resulting posterior probabilities as a function of c_α [for $P(x_1) = .5$]. Errors occur where the posterior is uncertain, near $P(x_1 \mid c_\alpha) = .5$. The steeper the posterior, the lower the chance of error; in turn, this happens when the likelihood distributions are well separated.

Next, we can write down the equivalent of the Q^* values from Equation 1 as the expected returns for each action, given the observation c_α (rather than the state x_1 or x_2):

$$Q^*_{c_\alpha}(L) = \mathbb{E}\left[r_{c_\alpha}(L) \mid c_\alpha \right] = P(x_1 \mid c_\alpha) r_1(L)$$
$$+ P(x_2 \mid c_\alpha) r_2(L) \qquad (12)$$

and

$$Q^*_{c_\alpha}(R) = \mathbb{E}\left[r_{c_\alpha}(R) \mid c_\alpha \right] = P(x_1 \mid c_\alpha) r_1(R)$$
$$+ P(x_2 \mid c_\alpha) r_2(R). \qquad (13)$$

These value expressions are functions of the cue c_α only through the belief state, which in this sense serves as a *sufficient statistic* for the cue in computing them. Another way of saying this is that the belief state satisfies the Markov independence property on which our reinforcement learning analysis relies: Given it, the future reward expectation is independent of the past (here, the cue). By determining the value expectations for each action, the belief state plays the role of the state from the previous section, which is unobservable here.

Given these values, then, as in Equations 2A and 2B, we can choose an optimal policy,

$$\pi^*_{c_\alpha} = \operatorname{argmax}_{a \in \{L, R\}} \left[Q^*_{c_\alpha}(a) \right], \qquad (14)$$

which, by direct calculation, turns out to just take the form of a threshold on the belief state or, equivalently, on the likelihood ratio $l(c_\alpha)$, and can be written as

$$\pi^*_{c_\alpha} = \begin{cases} L & \text{if } l(c_\alpha) > \theta, \\ R & \text{if } l(c_\alpha) < \theta, \\ \times & \text{if } l(c_\alpha) = \theta, \end{cases} \qquad (15)$$

where $\times$ implies that either L or R should be chosen with equal probability.

Given our assumption that $r_1(L) > r_1(R)$, the Bayes-optimal threshold θ_B is determined by the rewards and priors according to

$$\theta_B = \left(\frac{r_2(R) - r_2(L)}{r_1(L) - r_1(R)} \right) \left(\frac{P(x_2)}{P(x_1)} \right), \qquad (16)$$

which comes from the point at which the values of the two actions are equal—that is,
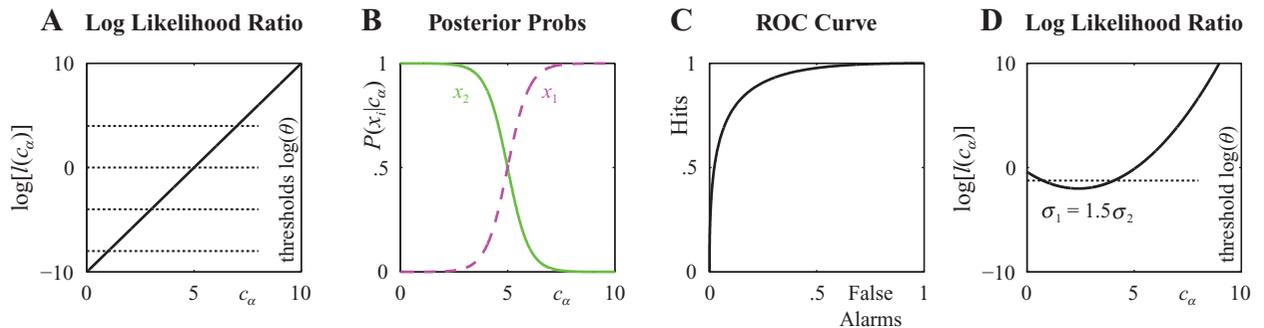
$$Q^*_{c_\alpha}(L) = Q^*_{c_\alpha}(R).$$

Figure 6. Signal detection theory for the Gaussian case. (A) The log likelihood ratio indicates how the observation c_α favors one or the other state x_i. For the equal-variance Gaussian case, this is linear in the observation c_α. Standard decision-theoretic tests are based on thresholding the likelihood ratio (or its logarithm, since this is a monotonic transform). (B) Bayesian decision theory is based on the posterior probabilities $P(x_i \mid c_\alpha)$, which combine the likelihood ratio and any prior information. Given the reward structure in Figure 5 and equal priors, these are also the Q^* values for the choices L and R. (C) The receiver-operating characteristic (ROC) plots the two independent quantities size (false alarm rate) and power (hit rate) of the test against each other. The area under the curve is related to the discriminability d' and is a measure of the quality of the cue c_α for distinguishing x_1 from x_2. (D) If $\sigma_1 \neq \sigma_2$, the log likelihood ratio need not be monotonic in c_α, and so implementing a single threshold on log $[l(c_\alpha)]$ can require more than one threshold on c_α.

To summarize, Bayesian decision theory in the case of state uncertainty is formally just the same as that in the case of complete knowledge of the state, except that the state is redefined to be the belief state, $P(x_\alpha = x_1 | c_\alpha)$. Given this, the same ideas as those in the previous section apply in terms of maximizing the expected return.

In standard decision theory, there is an important result called the *Neyman–Pearson lemma*, which implies that decisions should again be based on thresholds associated with the likelihood ratio $l(c_\alpha)$. However, unlike the Bayesian analysis that defines a single policy maximizing expected reward, misses and false alarms are not traded off against each other directly in standard decision theory, and so there is just a whole, one-dimensional family of optimal tests created by varying the threshold θ. The trade-off is depicted in the famous receiver-operating characteristic (ROC) curve, illustrated in Figure 6C, which plots the probability of a hit against that for a false alarm, across the whole range of thresholds. The area under the ROC curve is a measure of the quality of the cue c_α for discriminating x_1 and x_2, which itself is determined by the separation of the two likelihood distributions shown in Figure 5B. It is also related to other signal detection quantities, such as the discriminability, d' (Green & Swets, 1966).

Algorithmic Approaches

Here, we assume that subjects have knowledge of the rules (the conditional distributions, priors, and rewards) and must determine or learn the optimal policy. Model-based Bayesian methods are algorithmically simple, given this knowledge; they correspond literally to the derivation of the optimal policy outlined above. Model-free methods act to learn values such as $Q_{c_\alpha}^*(\mathrm{L})$ or, more directly, the best policy $\pi_{c_\alpha}^*$, without explicit reference to the distributions and priors; they must, instead, learn, as before, by bootstrapping from sampled experience. Under the same assumptions about exploitation as above—that is, that we are not trying to solve the exploration/exploitation problem—it is again sensible to consider the same learning rules.

The main issue that makes this different is that values and policies are functions of the continuous, real-valued quantity c_α—or of the continuous one-dimensional belief state that summarizes it—rather than of a discrete quantity like x_1 in the fully observable MDP. Furthermore, although the policy dependency actually takes a simple form—a single threshold θ—in terms of the belief state, model-free methods cannot directly compute the belief state and, so, face a more complicated problem. For instance, in the case in which $\sigma_1 = \sigma_2$, the optimal decision can also be described by a threshold in the observable quantity c_α. However, if $\sigma_1 \neq \sigma_2$, then, as is illustrated in Figure 6D, the likelihood ratio $l(c_\alpha)$ is not monotonic in c_α, and therefore two thresholds θ_l and θ_u are necessary, with

$$t(c_\alpha) = \begin{cases} \mathrm{L} & \text{if } c_\alpha < \theta_l \text{ or } c_\alpha > \theta_u, \\ \mathrm{R} & \text{if } \theta_l < c_\alpha < \theta_u, \\ \times & \text{if } c_\alpha = \theta_l \text{ or } c_\alpha = \theta_u. \end{cases} \quad (17)$$

Similarly, the nature of the dependence of values such as $Q_{c_\alpha}^*(\mathrm{L})$ on c_α is determined by quantities to which model-free methods have no direct access. One general solution is to use a flexible and general form for representing functions—for instance, writing

$$Q_{c_\alpha}^*(\mathrm{L}) = \sum_k f_k(c_\alpha) w_k(\mathrm{L}). \quad (18)$$

Here, $f_k(c_\alpha)$ are so-called *basis functions* of c_α, and w_k are parameters or weights whose settings determine the function. Depending on properties of $Q_{c_\alpha}^*(\mathrm{L})$ such as smoothness, a close approximation to it can result from relatively small numbers of basis functions. Furthermore, the model-free methods described in the previous section can be used to learn the weights.

Similarly, model-free and value-free policy gradient methods can be used to learn weights that parameterize a policy $\pi_{c_\alpha}^*$ directly. As is frequently the case, the policy (just one, or sometimes more, thresholds) may be much simpler than the values (a form of sigmoid function), making it potentially easier to learn appropriate weights.

Psychology and Neuroscience

The ample studies of human and animal psychophysics provide rich proof that subjects are sophisticated signal detectors and deciders in the terms established above. Behavior is exquisitely sensitive to alterations in the payoffs for different options (Stocker & Simoncelli, 2006) and changes in the observations (Körding & Wolpert, 2004); subjects even appear to have a good idea about the noise associated with their own sensations (Whiteley & Sahani, 2008) and actions (Trommershäuser, Landy, & Maloney, 2006; Trommershäuser, Maloney, & Landy, 2003a, 2003b) and can cope with even more sophisticated cases in which cues are two-dimensional (visual, c_α^v; auditory, c_α^a) and are conditionally independent given the state (Battaglia, Jacobs, & Aslin, 2003; Ernst & Banks, 2002; Jacobs, 1999; Yuille & Bülthoff, 1996).

There is also a range of influential neurophysiological investigations into the neural basis of these sorts of decisions. One set, prefigured in Figure 1B and executed by Britten et al. (1996; Britten et al., 1992) and Shadlen and Newsome (1996) and their colleagues, has focused on the processing of visual motion in area V5/MT in monkeys. Monkeys observe random-dot kinematograms, in which small dots appear and jump in various directions (Figure 7A). A proportion of the dots moves coherently, in the same direction, and the remainder moves at random; the job of the monkey is to report which direction the coherently moving dots favor on the basis of 2 sec worth of observing the motion. Typically, the monkeys have only two choices, 180° apart—that is, up and down in the figure. The filled-in circles in Figure 7B show the average performance of the monkey at this task as a function of the coherence level; for well-trained animals, performance is near perfect at a level of around 10%–15%. This is often called a *psychometric curve*.

Figure 7C shows example histograms of the firing rates of an MT neuron over the relevant period when faced with these stimuli, as a function of the coherence of the stimulus (i.e., the percentage of the random dots moving in a co-

A

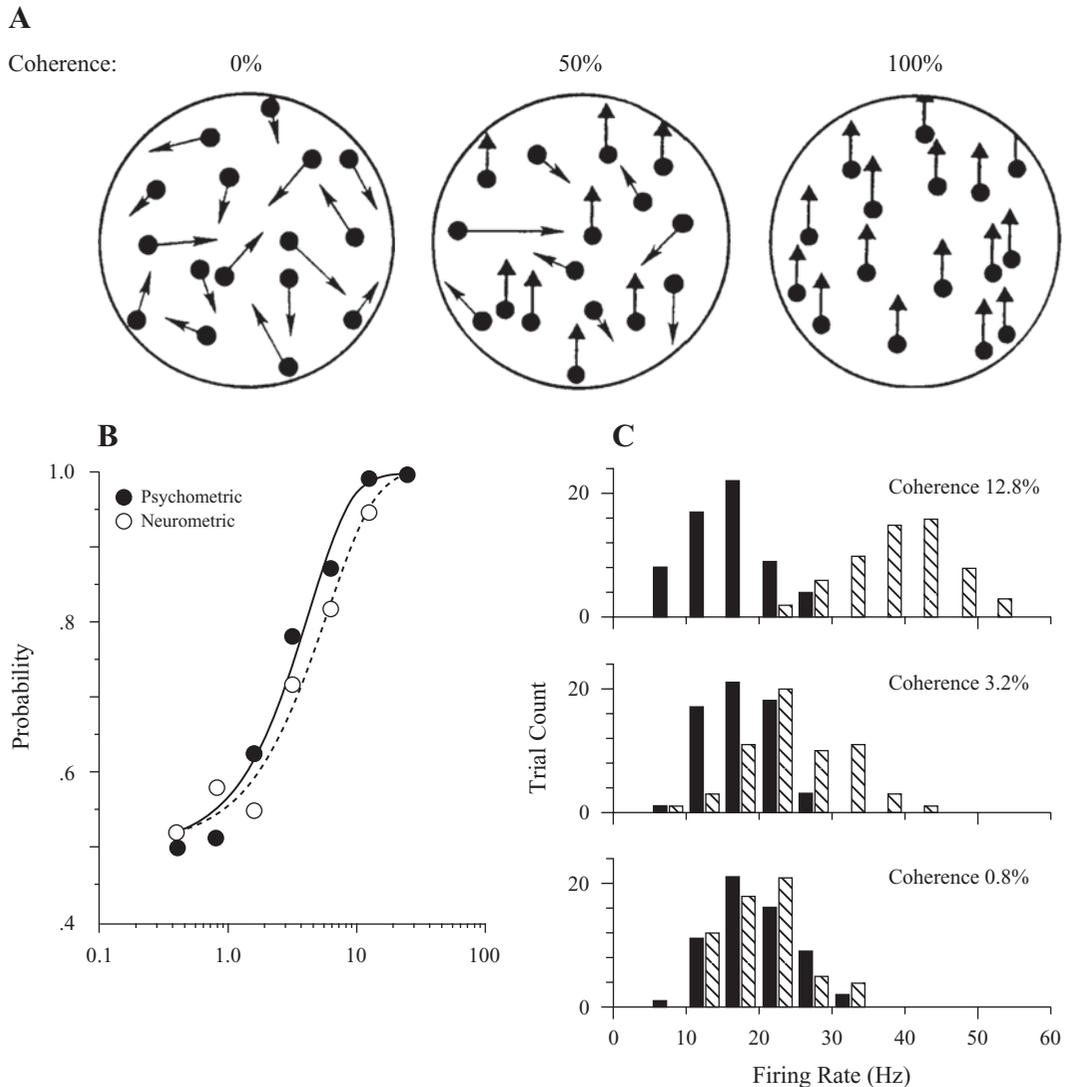Coherence: 0%   50%   100%

B

C

Figure 7. Britten, Shadlen, Newsome, and Movshon's (1992) experiment on primate signal detection. (A) Macaque monkeys observed random-dot motion displays made from a mixture of *coherent* dots interpreted as moving in one direction and *incoherent* dots moving at random. For these dots, the task was to tell whether the coherent collection was moving up or down. The percentage of coherently moving dots determined difficulty. (B) The filled points show the psychometric curve—that is, the discrimination performance as a function of the percentage of coherent dots. The open points show the quality of performance that would be optimally supported by a recorded neuron. (C) The graphs show histograms of the activity of a single MT neuron at three coherence levels over a 2-sec period, when the coherent motion was in its preferred direction (hashed) or opposite to this (solid). The larger the coherence, the larger the discriminability d', and the more easily an ideal observer counting the spikes just of this neuron would be able to judge the direction. From "The Analysis of Visual Motion: A Comparison of Neuronal and Psychophysical Performance," by K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon, 1992, *Journal of Neuroscience, 12*, pp. 4746, 4751, 4752, copyright 1992 by the Society for Neuroscience, and *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (pp. 89, 90), by P. Dayan and L. F. Abbott, 2005, Cambridge, MA: MIT Press, copyright 2005 by MIT. Adapted with permission.

herent direction), for both of the two directions of motion. Mapping this onto our problem, the firing rate is the cue c_α (from the perspective of neurons upstream), the state is the actual direction of motion of the stimulus, and the histograms in the figure are the conditional distributions $p_1(c_\alpha|x_1)$ (hashed) and $p_2(c_\alpha|x_2)$ (solid). It is apparent that these distributions are well separated for high-coherence trials—thus supporting low error discrimination—and less well for low-coherence ones.

The open circles in Figure 7B show the remarkable conclusion of this part of the study. These report the result of the Bayesian decision-theoretic analysis described above, applied to the neural activity data of the neuron shown in Figure 7C. This so-called *neurometric* curve shows the probability that an ideal observer knowing the firing rate distributions of a single cell and making optimal decisions would get the answer right. This single cell would already support decisions of the same quality as those made by the

whole monkey. Of course, the monkey's problem is to pick out the cells of this caliber (and particularly, collections of cells whose activity is as independent as possible, given the motion direction; Shadlen et al., 1996), integrate their activity over the duration of the trial, and limit the ability of noise to affect their actual decisions. The difficulty of doing these things should mitigate our surprise that the overall performance of the monkey is not substantially better than that of a single, somewhat randomly recorded neuron.

TEMPORAL STATE UNCERTAINTY

The examples of the last two sections can be combined to show how belief state estimation and reinforcement learning can be combined to find optimally exploitative decisions in POMDPs. This is exemplified by the other version of the task shown in Figure 1B, in which monkeys have to choose not only the direction of the motion, but also when they are sufficiently confident to make this choice. Here, they must balance the benefits of making their decision early—namely, avoiding the costs of waiting—against the change of making the wrong decision and getting no reward at all.

The Computational Problem

Figures 8A and 8B show a version of the task that combines some of the MDP aspects of the Markov Decision Problem section with the state uncertainty of the Signal De-

tection Theory section. Here, subjects start at $x_\beta \in \{x_3, x_4\}$ and again see a cue (referred to as c_β) that only partially distinguishes these states. They could choose L (which is correct at x_3), R (correct at x_4), or C, which incurs a small penalty (-0.1), but delivers them to $x_\alpha \in \{x_1, x_2\}$. Choosing C might be wise, if the cue available there c_α (assumed to be suitably independent of c_β) better resolves their state uncertainty (i.e., making them more sure about which of x_1 or x_2 they occupy than they were between x_3 and x_4), and, thus, more certain to get the reward for choosing L or R according to their beliefs. The choice of C is called *probing* and can be considered to be a form of exploration.

The Bayesian decision-theoretic ideas articulated in the Markov Decision Problem section extend smoothly to this case, just taking into account the idea that the subject's state should actually be its subjective belief state, given its observations. We first will consider the evolution of the belief state and then will see how this is employed to make optimal decisions. The graphs in Figure 9 refer to the Gaussian likelihood distributions used above and shown in Figure 8.

The case for x_3 and x_4 here is just the same as that for x_1 and x_2 in the Signal Detection Theory section. Given c_β, the posterior probability $P(x_3 | c_\beta)$ of being in x_3 is given by Bayes' rule, just as in Equation 10, proportional to the prior $P(x_3)$ and the likelihood $p_3(c_\beta | x_3)$.

Now by choosing C at the first stage, the subject will observe c_α at $x_\alpha \in \{x_1, x_2\}$; given this observation, it is again
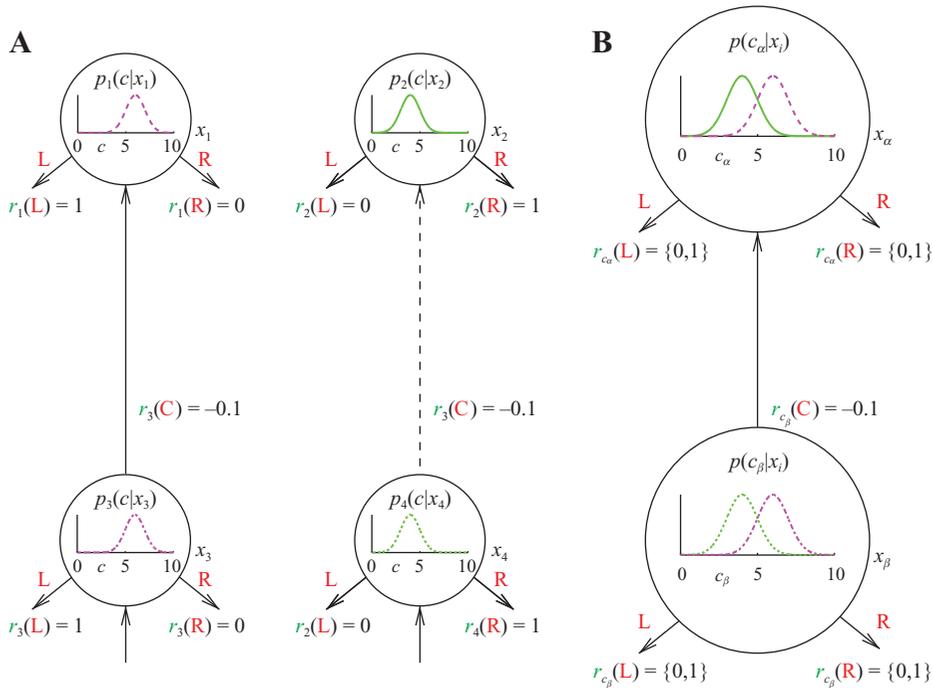

Figure 8. Information integration and probing. Subjects start at $x_\beta \in \{x_3, x_4\}$, but with uncertainty due to an aliased cue c_β, and can either act (perform L or R) immediately or perform action C, which incurs a small cost $r_{c_\beta} = -0.1$, but takes them to $x_\alpha \in \{x_1, x_2\}$, where a new, independent observation c_α can help resolve the uncertainty as to which of L or R would be better. As in Figure 5, panel A shows the objective state and outcomes; panel B shows similar quantities from a subjective viewpoint. The distributions show how the cues are related to the states. This is a simple task, since the transitions are deterministic $p_{31} = p_{42} = 1$.
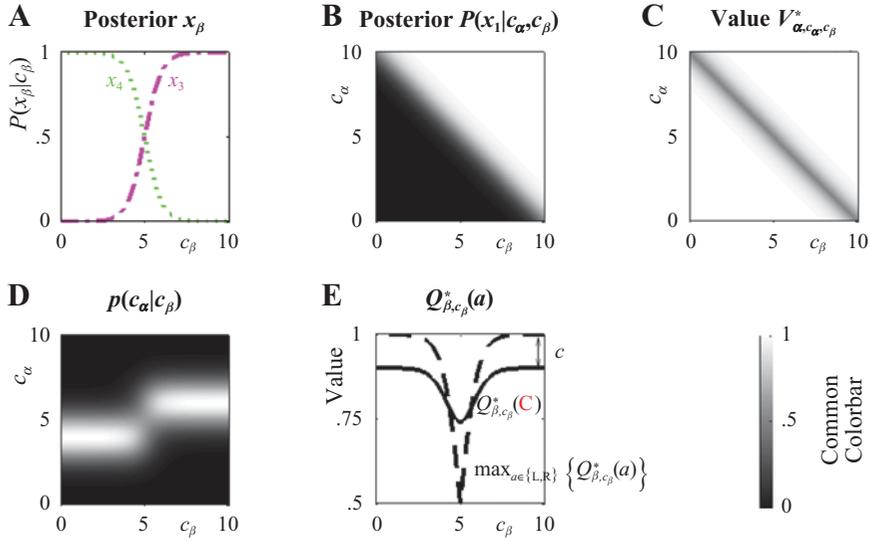
Figure 9. The value of sampling. (A) The posterior distribution $P(x_\beta = x_3 | c_\beta)$ at the first state x_β is just as in Figure 6C. **(B)** The log likelihood ratios just add, to give the posterior distribution $P(x_1 | c_\alpha, c_\beta)$. **(C)** The value of state x_α depends on c_α, c_β according to the maximum $\max\{P(x_1 | c_\alpha, c_\beta), P(x_2 | c_\alpha, c_\beta)\}$, since the subject will choose L or R according to these probabilities. **(D)** At x_β, the subject has to use c_β to work out the chance of seeing c_α at x_α. **(E)** Averaging the value in panel C over the distribution in panel D, and including the cost $r_{c_\beta}(C) = -0.1$, gives the value $Q^*_{\beta, c_\beta}(C)$ of probing (solid line). This is greater than the value of choosing the better of L and R (dashed line) for values of c_β that create the least certainty about x_β.

straightforward to compute a new belief state $P(x_1 | c_\beta, c_\alpha)$ using the previous belief state together with the the transition and cue probabilities. Because the successive observations are independent, and since the only way to get to x_1 is from x_3, the update takes a particularly simple form as the product of the previous belief state with the new observation probability:

$$P\left(x_1 | c_\beta, c_\alpha\right) \propto p_1\left(c_\alpha | x_1\right) P\left(x_3 | c_\beta\right)$$

$$= p_1\left(c_\alpha | x_1\right) p_3\left(c_\beta | x_3\right) P\left(x_3\right). \quad (19)$$

In short, the simple form of the problem means that incorporating each new cue into the belief state just involves multiplying the likelihood terms associated with the new observations (and renormalizing to make the sum of the beliefs equal to 1).

Note that the belief state after the second step depends on both cues, but it depends on the first cue only by way of the previous belief state $P(x_3 | c_\beta)$. Similarly, as it turns out, the expected future rewards will depend on both cues, but only through the belief state. This is an instance of a general and important fact about multistep problems with a hidden state. In general, cues like c_α will not suffice to determine future expected reward, because the entire previous history (in this case, just c_β) may still be relevant. However, in POMDPs, the current belief state is always a sufficient statistic for the entire cue history; that is, unlike any individual cue, it satisfies the Markov independence property. This is why it can always be used in place of an observable state for reinforcement learning.

To calculate the optimal policy, we now will proceed exactly as in the Markov Decision Problem section, *backward* from the second state to the first. If the subject gets to x_1 or x_2, it will be able to make a choice on the basis of the observation c_α. Given the rewards shown in Figure 8, it will choose L if $P(x_1 | c_\alpha, c_\beta) > P(x_2 | c_\alpha, c_\beta)$, and R otherwise. Thus, the value of being at x_α is

$$V^*_{\alpha, c_\alpha, c_\beta} = \max\left\{P\left(x_1 | c_\alpha, c_\beta\right), P\left(x_2 | c_\alpha, c_\beta\right)\right\}. \quad (20)$$

This is shown in Figure 9C. The value is high (white) when c_α and c_β are such that the subject can be rather sure whether $x_\alpha = x_1$ or $x_\alpha = x_2$—that is, when it can be sure which action to perform. If, instead, the cues are inconsistent, the value is closer to .5, which is that of random guessing.

However, the subject has to decide whether it is worth choosing C at state x_β *before* observing c_α. Thus, to work out the value of doing so, it has to average over what c_α might be, which in turn depends on the probability accorded to ending up in state x_1 before seeing c_α. Figure 9D shows the conditional distribution $p(c_\alpha | c_\beta)$; note, for instance, that if c_β strongly favors x_3, the subsequent state is likely to be x_1, so the distribution is close to $p_1(c_1 | x_1)$.

Averaging over this distribution, we get the mean value

$$V^*_{\alpha, c_\beta} = \mathbb{E}_{p\left(c_\alpha | c_\beta\right)}\left[V^*_{\alpha, c_\alpha, c_\beta}\right] \quad (21)$$

$$= P\left(x_3 | c_\beta\right) \int_{c_\alpha} V^*_{\alpha, c_\alpha, c_\beta} p\left(c_\alpha | x_1\right) dc_\alpha$$

$$+ P\left(x_4 | c_\beta\right) \int_{c_\alpha} V^*_{\alpha, c_\alpha, c_\beta} p\left(c_\alpha | x_2\right) dc_\alpha, \quad (22)$$

since the transitions can occur only from $x_3 \rightarrow x_1$ or $x_4 \rightarrow x_2$. This is the exact equivalent of Equations 3A and 3B, except that taking an expectation over the belief state requires an integral (because it is continuous) rather than a discrete sum.

The last step is to work out the values of executing each action at x_β. For action C, from Equation 4 and the transition probabilities:

$$Q^*_{\beta,c_\beta}(C) = r_{c_\beta}(C) + V^*_{\alpha,c_\beta} = -0.1 + V^*_{\alpha,c_\beta}. \qquad (23)$$

This value is shown as the solid line in Figure 9E. It is high when the subject can expect to be relatively certain about the identity of x_α and low when this is not likely. The maximum value is $1 + r_{c_\beta}(C) = 0.9$, given the cost -0.1 of probing.

The expected values to the subject of performing L or R at x_β are

$$Q^*_{\beta,c_\beta}(L) = P\left(x_3 \mid c_\beta\right) \qquad Q^*_{\beta,c_\beta}(R) = P\left(x_4 \mid c_\beta\right), \quad (24)$$

which are the exact analogues of the $Q^*_{c_\alpha}(a)$ terms in Equation 12. The dashed line in Figure 9E shows the value of the better of L and R. This is near 0.5 for intermediate values of c_β, where the subject will be very unsure between x_3 and x_4 and, thus, between the actions. In this region, action C is preferable because the additional observation c_α is likely to provide additional certainty and a better choice at the next step, even weighed against the -0.1 cost of C.

Combining the equations above, the subject should choose C if

$$\max\left\{P\left(x_3 \mid c_\beta\right), P\left(x_4 \mid c_\beta\right)\right\} <$$

$$-0.1 + \mathbb{E}_{p\left(c_\alpha \mid c_\beta\right)}\left[\max\left\{P\left(x_1 \mid c_\alpha, c_\beta\right), P\left(x_2 \mid c_\alpha, c_\beta\right)\right\}\right];$$
$$(25)$$

that is, if the benefit of the added certainty about being in x_1 or x_2 outweighs the cost -0.1 of sampling.

The two points at which the two curves in Figure 9E cross are thresholds c_l and c_h in c_β, such that probing is preferred when $c_l < c_\beta < c_h$. We will see that this sort of test is quite general for problems with this character, although the thresholds are normally applied to the belief states associated with the observations, rather than to the observations themselves.

Algorithmic Issues

The case of temporally extended choices in the face of incomplete knowledge is known to pose severe computational complications as the number of states grows larger than that for the simple problem described here. The difficulties have to do with the definition of the state of the subject. There are actually two ways to look at this, both of which are problematic. One way is to see the state as a summary of the entire history (and, for planning, the future) of the observations (say, $c_\beta, c_\alpha, \ldots$) of the agent. Indeed, we indexed value functions such as $V^*_{\alpha,c_\alpha,c_\beta}$ by this history. The trouble with this representation is that the his-

tory grows over time (the number of steps in the problem might be much larger than the two here) and so the dimensionality of the optimization problem grows also. It also poses severe demands on short-term memory. In general, it is not possible to represent optimal value functions and policies with only a few basis functions, as in Equation 18.

The alternative representation, which we have stressed, is to note that, given a model, the full history of observations can be summarized in a single belief state, $P(x_\tau \mid \{c_1, \ldots, c_\tau\})$, which can then be updated recursively at each step. This has the advantage of not changing dimension over time and, thus, also not placing such an obvious load on working memory. However, like the cue history (but unlike the states of an observable MDP), this probability distribution is still a multidimensional, continuous object, which makes learning values and policies as functions of it still difficult in general.

Since belief states of this sort are computed by inference using a model, model-free methods cannot create them and, therefore, generally have to work with the history-based representation. However, the Markov sufficiency of the belief state immediately suggests an appealing hybrid of model-free and model-based approaches, whereby a model might be used only to infer the current belief state and then model-free methods are used to learn values or policies on the basis of it (Chrisman, 1992; Daw, Courville, & Touretzky, 2006). This view separates the problem of state representation from that of policy learning: The use of a model for state inference addresses the insufficiency of the immediate cues c. Having done so, it may nevertheless be advantageous to use computationally simple model-free methods (rather than laborious model-based dynamic programming) to obtain values or policies.

When belief states are used, algorithmic issues also arise in updating them using Equation 19. Notably, it may be simpler to represent the belief state by its logarithm, in which case the multiplication to integrate each new observation becomes just a sum. The idea of manipulating probabilities in the log domain is ubiquitous in models of the neural basis of this sort of reasoning (Ma, Beck, Latham, & Pouget, 2006; Rao, 2004), including the one discussed next. However, it is important to note that in general, belief updates involve not just multiplication, as above, but also addition, which means that the expression is no longer simplified by a logarithm. For instance, if it were possible to get to state x_1 from both x_3 and x_4, Equation 19 would sum over both possibilities:

$$P\left(x_1 \mid c_\beta, c_\alpha\right) \propto p_1\left(c_\alpha \mid x_1\right) P\left(x_1 \mid x_3\right) P\left(x_3 \mid c_\beta\right)$$

$$+ p_1\left(c_\alpha \mid x_1\right) P\left(x_1 \mid x_4\right) P\left(x_4 \mid c_\beta\right). \quad (26)$$

Purely additive accumulation is therefore limited to a class of problems with constrained transition matrices and independent observations. Also, it is only in a two-alternative case that the normalization may, in general, be eliminated by tracking the ratio, or log ratio, of the probability of the states, as in Equation 11.

Finally, in discussing partially observable situations, we have so far assumed that the model is known. In general, of course, this might also be learned from experience. In the observable MDP, this simply requires counting state transitions (Equation 6B); however, when states are not directly observable, their transitions obviously cannot be counted. One family of algorithms for learning a model in the face of hidden states involves so-called expectation maximization methods (Dempster, Laird, & Rubin, 1977). The algorithm takes a starting guess at a model and improves it, using two steps. In the first, the *expectation* phase, the model is assumed to be correct, and the inference algorithms we have already discussed are used to infer which (hidden) state trajectory was responsible for observed cues. In the second, *maximization* phase, these beliefs about the hidden trajectory are treated as being correct, and then the best model to account for them is inferred using a counting process analogous to Equations 6A and 6B. These phases are repeatedly alternated, and it can be shown that each iteration improves the model, in the sense of making the observed data more likely under it (Neal & Hinton, 1998). In actuality, these model-learning methods require substantial data, are difficult to extend to practical online learning from an ongoing sequence of experience, and are prone to getting stuck at suboptimal models that occupy local maxima of the hill-climbing update.

Psychology and Neuroscience: Drift Diffusion Decision Making

In the original studies described in Figure 1B and Figure 7, which linked the activity of MT neurons to choice behavior, the monkeys observed the random dot motion displays for a fixed period of 2 sec before making their choice as to its direction of motion. The other version of the task, in which the the subjects are free to choose *when* to make their decision, has been the topic here. It is a version of one of the most important developments in decision theory—namely, the SPRT (Gold & Shadlen, 2001, 2007; Ratcliff & Rouder, 1998; Shadlen et al., 2007; Smith & Ratcliff, 2004; Wald, 1947), which is a highly active area of investigation in both psychology and neuroscience.

The SPRT is designed for the circumstance in which subjects receive a stream of observations (like c_β, c_α, but continuing potentially indefinitely: . . . , c_τ, . . .), pertaining to a binary discrimination (in our case, between x_3, x_1, both of which require one choice, and x_4, x_2, which require a different choice). We will call all the states that require L x_L, and all those that require R x_R. Subjects can choose at any time (picking L or R), or they can wait (C) and sample more information. In our framework, we would seek Bayesian optimal choices in the light of the costs; the SPRT is derived from the slightly different goal of minimizing the average decision time for a given probability of getting the answer correct.

As in all of our examples, the critical observation is that the subjective state of the subject at stage T, which comprises the information required to calculate the current posterior over the choices at that stage, depends only on the belief state or something equivalent to it. For the SPRT, this is normally represented as the accumulated log likelihood

ratios of all the pieces of evidence $\mathbf{c}_T = \{c_1, \ldots, c_T\}$ given the two state possibilities, which can be written

$$
\begin{aligned}
\ell_T(\mathbf{c}_T) &= \log \frac{p(c_1, \ldots, c_T \mid x_L)}{p(c_1, \ldots, c_T \mid x_R)} \\
&= \sum_\tau \log \frac{p(c_\tau \mid x_L)}{p(c_\tau \mid x_R)} \\
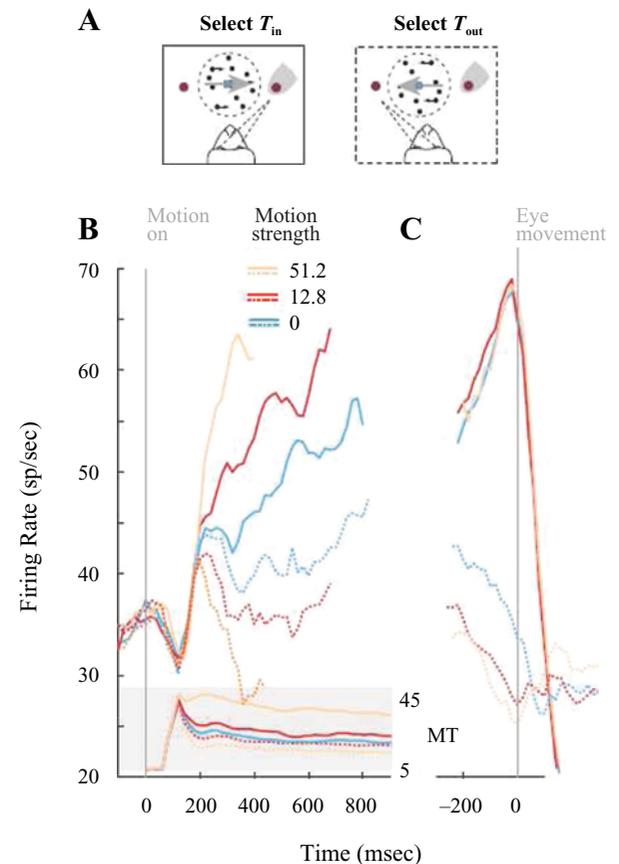&= \sum_\tau \ell_\tau(c_\tau),
\end{aligned}
$$

Figure 10. Evidence accumulation in the lateral intraparietal area (LIP). This figure is taken from Gold and Shadlen (2007) and is based on data from Britten, Shadlen, Newsome, and Movshon (1992) and Roitman and Shadlen (2002) and shows the putative integration of opponent evidence from MT to construct a net log likelihood ratio associated with an SPRT-like threshold-based decision. (A) Monkeys discriminate the direction of motion and respond by making saccades either into (T_{in}) or away from (T_{out}) the response field of LIP neurons. (B) The main plots show the average activity of LIP neurons as a function of motion strength or coherence, temporally triggered on the motion onset. Solid lines are T_{in} cases; dashed lines, T_{out} cases. Following a transient, the curves evolve in a manner roughly consistent with a putative log likelihood ratio. The inset plot shows the constancy of the activity of MT neurons, albeit from a different experiment. (C) These plots are triggered on the time of the saccade, to examine a threshold-like policy. From "The Neural Basis of Decision Making," by J. I. Gold and M. N. Shadlen, 2007, *Annual Review of Neuroscience, 30*, p. 548. Copyright 2007 by Annual Reviews. Reprinted with permission.

given independent evidence and the trivial transition structure. Thus, a decision maker need only keep track of this running quantity, plus, at most, some function of the index of the current stage T. The SPRT is an extremely simple test that uses two thresholds, ϕ_l and ϕ_h, and has

$$
\pi\left(c_1 \ldots c_T\right)=\begin{cases}\text{L} & \text{if } \ell_T\left(\mathbf{c}_T\right) \geq \phi_h \\ \text{R} & \text{if } \ell_T\left(\mathbf{c}_T\right) \leq \phi_l \\ \text{C} & \text{if } \phi_l < \ell_T\left(\mathbf{c}_T\right) < \phi_h\end{cases} \quad (27)
$$

or the equivalent for other representations of the belief state, which are typically monotonic in the log likelihood ratio.

We saw, in discussing Figure 9E, that the test for performing C for the problem in Figures 8A and 8B has exactly this form too. The surprising fact about the SPRT, which follows from the Markov property together with a constraint that the cues at each stage be independent and identically distributed, is that the thresholds ϕ_l and ϕ_h are independent of the stage or time.

Implementing tests such as the SPRT or other tests associated with the basic task would, therefore, require *integration* of the belief state associated with the observations, plus a decision rule that will look like a threshold on the belief state.

Figure 10 shows schematic results of an experiment (shown as a cartoon in panel A) in which animals had a free choice of decision time, reporting their decision by making a saccade from a fixation point to one of two targets (Gold & Shadlen, 2001, 2007; Roitman & Shadlen, 2002). The main curves in Figures 10B and 10C show the average activity of neurons in the lateral intraparietal area (LIP), a site that may report the output of the putative integrator. Neurons in this area have eye movement response fields (the gray patches in Figure 10A); that is, they fire selectively when subjects are planning saccades to targets in a space of the sort employed in this experiment. The curves in Figures 10B and 10C show the mean firing rates of such neurons under various conditions over the course of trials.

Consider first the bottom inset to Figure 10B, taken from data in Britten et al. (1992), which shows the activity of MT neurons to random-dot motion (albeit in the fixed duration task) over the course of trials, locked to the motion onset for trials of different coherences (colors) in two opposite directions (solid or dotted). This is effectively a different way of looking at the data in Figure 7 and shows that following a transient response to the onset, their firing rates are rather constant.

Next, consider the solid curves in the main part of Figure 10B. These show cases in which the response was to the target defined by the LIP neurons (T_{in} cases, as in Figure 10A), for three different coherence levels (motion strengths). For the most coherent motion (highest; gold), following a (different) transient, this rises steeply. For less coherent motion (middle; red), this rises less steeply; for incoherent motion (lowest; blue), it rises still less steeply. The idea is that the LIP neurons report belief states for the choice associated with their response fields, perhaps represented as a log likelihood ratio computed by integrating opponent activity from MT neurons with opposing prefer-

ences. The belief states rise more steeply when the motion is stronger. For 0% coherent motion, they rise only because it is the trials on which the monkey ultimately chooses the particular response field that are averaged, and so these should be the ones in which the evidence, by chance, ultimately favors this direction.

Figure 10C shows the same conditions, but now triggered on the saccades themselves. All the curves lie on top of each other, even for the 0% coherence case, coming near to a single point just before the saccade happens. This is exactly the behavior to be expected from the action of an upper threshold (like ϕ_h in Equation 27), which is applied directly to the firing rate, giving rise to a response (after a short delay). The same considerations as those in the Signal Detection Theory section make it reasonable that the (trial-average) firing rate of single neurons could seemingly be directly associated with a response-triggering threshold.

The dashed lines in these figures report the activity averaged over cases in which the monkeys chose the *opposite* direction (T_{out}). The simplest version of the idea is that the cells continue to report the accumulated belief state, which now tends to decrease rather than increase. However, responses are not triggered by crossing a lower threshold ϕ_l; rather, there is an upper threshold ϕ_h for other neurons with response fields preferring the actual saccade location that is ultimately chosen.

The SPRT models cases in which information accumulates over time from an external source. However, it and related models have been used by a wealth of very important studies (e.g., Ratcliff & McKoon, 2008; Smith & Ratcliff, 2004) to characterize reaction times in cases such as search tasks, in which the external information is constant but *internal* processing associated with this information might unfold over time. There are also a number of influential connectionist (Usher & McClelland, 2001) and neural (Lo & Wang, 2006; Wang, 2002, 2006) models of this, together with mathematical theory about their interrelationship (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006).

There is an active theoretical and experimental debate about the nature of coding and decision making associated with LIP in tasks that are more complicated than this, involving coherences that vary from trial to trial or differential rewards or priors for different options and nonconstant thresholds (e.g., Glimcher, 2004; Platt & Glimcher, 1999; Yu, 2007). One trouble with the random-dot motion is that it is not clear what likelihood contributions to each direction of motion should arise from each segment of the stimulus; that is, the cues c_τ from our idealization are internal to MT and, therefore, not well controlled experimentally. A test of LIP's report of an integration process with more discretely controlled cues showed promising but partial results (Yang & Shadlen, 2007), with extreme values of the summed log likelihood ratios failing to be quite correctly represented.

The SPRT is a seminal contribution to the theory of decision making, and indeed, there is a subfield studying analytical methods rather different from the ones we have presented here for solving it (the basic results are beautifully reviewed in Shadlen et al., 2007). However, the

SPRT is brittle, in that most changes to the task will break its guarantees. In fact, even the case of more than two options is surprisingly complicated. This is one reason why here, we have framed this problem in terms of a more general model of decision making in the face of uncertainty, which gives a similar account of this particular task but readily accommodates a host of elaborations. However, as has already been mentioned, the solutions can be extremely hard to compute or even approximate.

Viewing the random-dots task as a partially observable reinforcement learning problem also suggests broadly how the brain might solve the policy-learning problem for it (Ahmadi & Latham, personal communication; Gold & Shadlen, 2002): by using general reinforcement learning techniques—such as those described in the Markov Decision Problem section and putatively implemented by systems such as midbrain dopamine—over a belief state representation (as described here for LIP). In other work, the idea that the dopamine system might learn by employing a belief state representation has also been used to explain how it might cope with partial observability arising from the unpredictable timing of cues and to explain some characteristics of dopaminergic responses in such situations (Daw, Courville, et al., 2006).

EXPLORATION AND EXPLOITATION

So far, we have considered exploitation: choices that maximize the expected single-episode return, given whatever is known about the rules by the time of that episode. As we have already mentioned, a more ambitious subject might wish to maximize *lifetime* utility, earning as much reward as possible over a whole series of episodes. This is exactly the goal in the task shown in Figure 1C.

Doing so requires balancing exploitation against *exploration*: taking choices that might not be expected to pay off as much immediately but, by improving knowledge of the rules, might improve the prospects for earning reward on subsequent trials. This is a difficult balance, but its elegant decision-theoretic solution follows directly from the analysis we have already developed. In particular, ignorance of the rules can be treated in exactly the same way that we treated ignorance of the state: by planning on the basis of beliefs, rather than directly on the basis of observations. In turn, the value of exploration can be quantified just as we have previously evaluated information-gathering (*probing*) actions like C.

The Computational Issue

Consider the one-state (x_1) example shown in Figure 11A, known in this context as a two-armed bandit. This time, the rewards are binary (0 or 1), and a choice of action L or R delivers reward 1 with probability p_L or p_R, respectively. The agent starts out ignorant of these probabilities, has a limited number (N) of trials (say, 50) in which to play the game, and aims to maximize the total obtained reward. The dotted lines emphasize the iterative nature of the task. The explore/exploit dilemma arises here because p_L and p_R remain the same throughout the 50 trials; choices not only earn immediate rewards, but also help the subject to learn the values of these probabilities, potentially improving subsequent choices.

If the subject starts with the additional prior knowledge that p_L and p_R were each drawn uniformly and independently between 0 and 1, clearly the expected value of either is .5 per choice, and the expected cumulative reward of blindly choosing either 50 times is 25. However, given two options whose payoffs were chosen in this manner, the expected value of the *better* one (whichever it is) is actually higher—namely, 2/3. If, therefore, the subject was *told* which one is better, he or she could exploit it and expect to earn about 33 rewards. Without this knowledge, but choosing so as optimally to balance exploring to find it and exploiting it, the subject can expect still to obtain about 96% of that value.
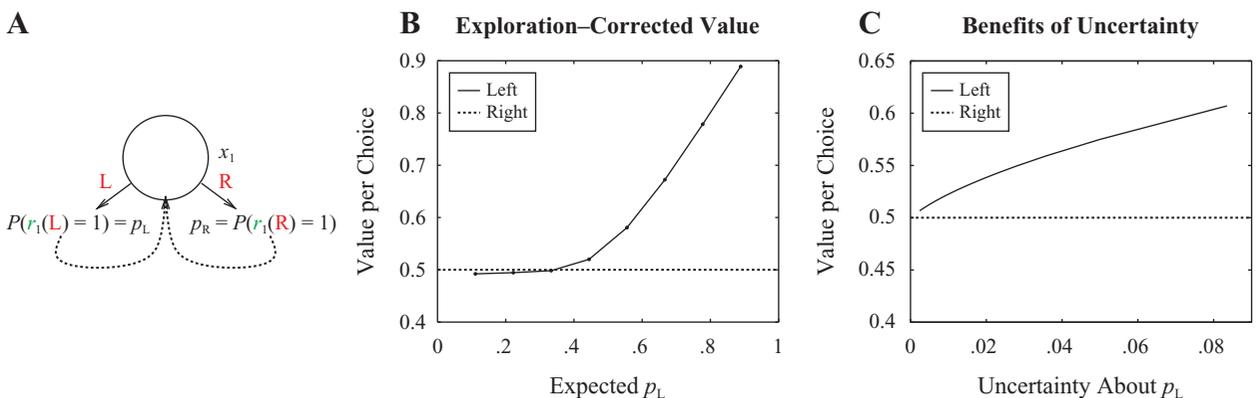
A

B   **Exploration–Corrected Value**

C   **Benefits of Uncertainty**

Figure 11. (A) Two-armed bandit task. For each episode, the agent can choose L or R, which pay off with (initially unknown) probabilities p_L and p_R. Subjects have a fixed total number of choices N (emphasized by the dotted lines) and seek to optimize their summed reward. (B) The value of exploration: The long-term value of choosing uncertain option L as a function of its expected immediate payoff. (The cumulative Q value over 50 choices is plotted normalized to a per choice value.) Note that even when the immediate payoff is expected to be less than 50%, the value lies above that for repeated choice of option R, which pays off 50% for certain. (C) The value of uncertainty: The long-term value of choosing uncertain option L as a function of how uncertain it is (measured as posterior variance), holding the expected immediate payoff fixed at 50%.

We can define this optimal balance, and compute its expected value, using dynamic programming in the space of belief states. This proceeds just as in the previous sections, except that the belief is over the rules (p_R and p_L), rather than over the state x_β. What we arrive at is a form of master policy specifying which option to choose, given any current beliefs about the rules.

In this task, p_R and p_L can be estimated at any stage, using counts of the number of times each option was rewarded and unrewarded (e.g., r_L and u_L for rewarded and unrewarded choices on the left). Since the prior distribution over p_L is uniform, the posterior distribution (the equivalent of that in Equation 10) is known as a beta distribution, with parameters $r_L + 1$ and $u_L + 1$. Various properties flow straight from this, such as the posterior mean, which is one estimate for p_L:

$$\mathbb{E}\left[p_L \right] = \frac{r_L + 1}{r_L + u_L + 2}. \qquad (28)$$

Importantly, we can denote any belief state by a vector of counts, $\langle r_L, u_L, r_R, u_R \rangle$; together, these counts are sufficient statistics for the entire history of the game.

Let us, then, consider the value $Q_{\langle 0,0,0,0 \rangle}(L)$ of choosing L at the start of the game. Either the choice is rewarded, in which case the new belief state will be $\langle 1,0,0,0 \rangle$, or it is not rewarded, in which case the belief state will be $\langle 0,1,0,0 \rangle$. Crucially, we know the probability of attaining either result: It is just that given by Equation 28—that is, 50%. That is, Equation 28 does not just give the subject's *best guess* as to the rules: Since it arises from a correct inference, given a presumptively true prior about how the rules were generated, this is also the *actual* probability, over games, that a choice will be rewarded, conditional on what the subject has observed so far. Therefore, working toward the same sort of recursion as that in Equation 4 by using the optimal values $V^*_{\langle 1,0,0,0 \rangle}, V^*_{\langle 0,1,0,0 \rangle}$ of the belief states consequent on either option, we can write

$$Q^*_{\langle 0,0,0,0 \rangle}\left(L\right) = 0.5 \cdot \left(1 + V^*_{\langle 1,0,0,0 \rangle}\right) + 0.5 \cdot \left(0 + V^*_{\langle 0,1,0,0 \rangle}\right) (29)$$

or, more generally,

$$Q^*_{\langle r_L, u_L, r_R, u_R \rangle}\left(L\right) = \frac{r_L + 1}{r_L + u_L + 2} \cdot \left(1 + V^*_{\langle r_L+1, u_L, r_R, u_R \rangle}\right)$$
$$+ \frac{u_L + 1}{r_L + u_L + 2} \cdot \left(0 + V^*_{\langle r_L, u_L+1, r_R, u_R \rangle}\right),$$

and similarly for the value of R.

Finally, just as in Equations 3A and 3B, the value V^* of a belief state is just that of the better choice there,

$$V^*_{\langle r_L, u_L, r_R, u_R \rangle} = \max\left[Q^*_{\langle r_L, u_L, r_R, u_R \rangle}\left(L\right), Q^*_{\langle r_L, u_L, r_R, u_R \rangle}\left(R\right) \right],$$

and we complete the recursive definition by defining the future values Q^* and V^* as zero at the end of the game, when no further choices remain.

Together, these equations quantify the explore/exploit trade-off. First, because of the boundary condition, at the

last choice (when $r_L + u_L + r_R + u_R = N - 1$), the value of each action is just its chance of immediate reward, given by Equation 28. Here, the subject should simply *exploit* by choosing the option with the better immediate reward expectation.

In contrast, further from the end of the game, exploration has value. To see why, consider a game in which p_R is known to be exactly .5 and we focus only on learning p_L. At the start of the game, the expected immediate payoff of .5 is the same for both L and R. However, the uncertainty surrounding this value for action L represents an opportunity: If the option's true reward probability turns out to be more than .5, finding this out will allow the agent to choose it, earning more. On the other hand, if it turns out to be less, the agent can always just go back to choosing action R. Therefore, even though the immediate expected return for L is the same as that for R, the information gained by choosing it gives the uncertain option a higher expected future return. This can be seen by considering the future value terms V^* in Equation 29: The future value $V^*_{\langle 1,0,0,0 \rangle}$ after a win on L will involve another left choice now expected to pay off with probability 2/3 (from Equation 28 with $r_L = 1$, $u_L = 0$); but even after losing on L, the future value $V^*_{\langle 0,1,0,0 \rangle}$ cannot be worse than that coming from the choice of the safe option R.

The extra future value from exploration means that it can be, on balance, worth choosing an option that is more uncertain, even if it has a *lower* immediate reward expectancy than does the alternative. Figure 11B illustrates this point: It shows the total expected values (over $N = 50$ trials; normalized per trial) of the first choice of a range of uncertain options L with different expected means that were generated by giving the agent seven observations of L prior to the game, in different mixtures of wins and losses. The dotted line shows the value of choosing the known 50% reference R.

Figure 11C illustrates directly how the value of exploration follows from uncertainty, by plotting how the value of choosing L increases monotonically as the uncertainty about its payoff probability increases, while holding the overall expected chance that it will pay off fixed at 50%. Here, the agent is given extra observations of equal numbers of wins and losses on L prior to starting the game; uncertainty is increased by reducing these numbers.

In short, exploration is valuable because it has the possibility of improving choices on future episodes, earning more reward in the long run. This is exactly the same reason why probing to reduce state uncertainty was worthwhile in the example in the previous section. In principle, the same analysis extends directly to exploration in unknown multistep decision processes such as those considered in the previous section, in which case the belief state includes beliefs about the transition probabilities, as well as the rewards. A different form of probing arises in this case, in that actions can be taken to reach areas of the state space that are poorly explored in order to learn about them. In practice, the exact solution to even the smallest such problems is intractable, due to the continuous nature or the high dimensionality of the belief state.

Algorithmic Issues

So far, we have characterized the computational issues underlying exploration as concerning ambition—that is, optimizing reward accumulated across episodes, rather than myopically exploiting within each. In this respect, the problem can be seen to relate to issues of temporal discounting: An unambitious subject is like an impatient one, who discounts future reward sharply. In practice, however, even for minimally patient subjects, the limiting constraint on exploratory behavior is more likely to be the extreme complexity of decision-theoretically optimal exploration. Rather than differences in discount preferences or goals, different exploratory behaviors can arise from different algorithmic approaches to the decision problem. For instance, ignoring uncertainty about the rules is what enables many of the standard reinforcement learning algorithms discussed in the Markov Decision Problem section to work; but this can just as easily be viewed as a simplifying assumption that precludes optimal exploration, rather than as a myopic goal.

In fact, even though the decision-theoretic analysis of exploration is formally equivalent to that of partial observability——indeed, leading to similar practical problems in actually solving for the optimal policy—researchers have developed a number of special purpose approaches to the exploration problem.

First, Gittins (1989) showed how a subset of exploration problems could be simplified by computing an *index*, similar to a *Q* value, for each action separately in a smaller (although still nontrivial) subproblem. The approach takes advantage of a sort of independence between the actions arising from the problem structure. It works for multiarmed bandit problems of the sort described above, although only when the number of trials, *N*, is either infinite (and future reward is discounted exponentially) or unknown with a constant hazard function—that is, a constant probability per trial of terminating.

However, most problems, including exploration in multistep decision problems, cannot be simplified this way. In reinforcement learning, heuristic approaches are, therefore, common, more or less qualitatively inspired by the optimal analysis above. Some approaches evaluate actions according to their expected exploitative value plus an *uncertainty bonus* intended to capture the additional value of exploration (Dayan & Sejnowski, 1996; Dearden, Friedman, & Russell, 1998; Kaelbling, 1993; Sutton, 1990). Authors differ as to how uncertainty should be estimated (which is itself not simple in this context) and the exact form of bonus itself. One particularly simple variation is the novelty bonus (e.g., Ng, Harada, & Russell, 1999), which approximates an uncertainty bonus for unexplored options simply by initializing estimates of their values "optimistically" high, encouraging their exploration until their true value is discovered.

A more primitive alternative, which is less well grounded in the analysis above, is to encourage exploration more blindly by introducing some sort of randomness into the choice process.

One aspect of all of these approximate approaches is that they typically require careful adjustment of such fac-

tors as the degree of randomness or the lucrativity of the uncertainty bonuses in order to perform well. Although finding the optimal exploratory policy is not normally considered a learning problem (since the goal, after all, is lifetime optimality), learning of a sort is implicated in tuning the parameters of the approximate approaches to improve their performance.

Psychology and Neuroscience

Outside the field of ethology (e.g., Mangel & Clark, 1989; McNamara & Houston, 1980; Pyke, 1984) exemplified by some early work on the exploration of birds (Krebs, Kacelnik, & Taylor, 1978), there is relatively little direct evidence as to how animals or humans explore. There are, however, a number of models. In this vein, Kakade and Dayan (2002) noted that dopamine neurons respond to novel but affectively neutral stimuli with a burst–pause response; they suggested that this pattern could be explained if the neurons were reporting the effect on reward prediction error of stimuli that were up-valued due to a novelty or uncertainty bonus.

Subsequently, in the task shown in Figure 1C, Daw, O'Doherty, et al. (2006) sought to test a similar idea more directly, in an fMRI study of humans making choices for money in a four-armed bandit task similar to the two-armed bandit task analyzed above. Subjects had additional uncertainty and pressure to explore, because the bandits' values were constantly changing. The authors sought to quantify the effect of this uncertainty on exploratory choices by fitting subjects' trial-by-trial behavior with a reinforcement learning model incorporating uncertainty bonuses but found no such influence. Instead, the subjects' exploration was best explained by a random exploration model, the so-called *softmax*, which, for two choices, involves a variant on Equation 8 with

$$P_1\left(a = \mathrm{L}\right) = \sigma\left\{\beta\left[Q_{\langle r_\mathrm{L}, u_\mathrm{L}, r_\mathrm{R}, u_\mathrm{R}\rangle}\left(L\right) - Q_{\langle r_\mathrm{L}, u_\mathrm{L}, r_\mathrm{R}, u_\mathrm{R}\rangle}\left(R\right)\right]\right\},$$

where $\beta > 0$ regulates the strength of competition—that is, the extent to which a small difference in expected values of the actions translates into a large difference in their probability of choice. This parameter is therefore another of the more primitive ways of trading off exploration and exploitation.

The neural data supported a similar conclusion. Bonus models such as that of Kakade and Dayan (2002) (or indeed, the optimal decision theoretic analysis) quantify the value of exploration and exploitation in terms of a single currency of expected future reward. They therefore predict that exploratory value should be reported just as is exploitation—for instance, through dopaminergic spiking. Contrary to this expectation, Daw, O'Doherty, et al. (2006) found that although dopaminergic efferent areas such as the striatum were activated by predictions and prediction errors for money, they exhibited no additional activation that might reflect bonus values involved in exploration. Instead, separate areas—the frontal pole and, also, an area of the intraparietal cortex—were activated preferentially when subjects explored rather than exploited (Figure 12). Here, exploration was operationally defined as a choice
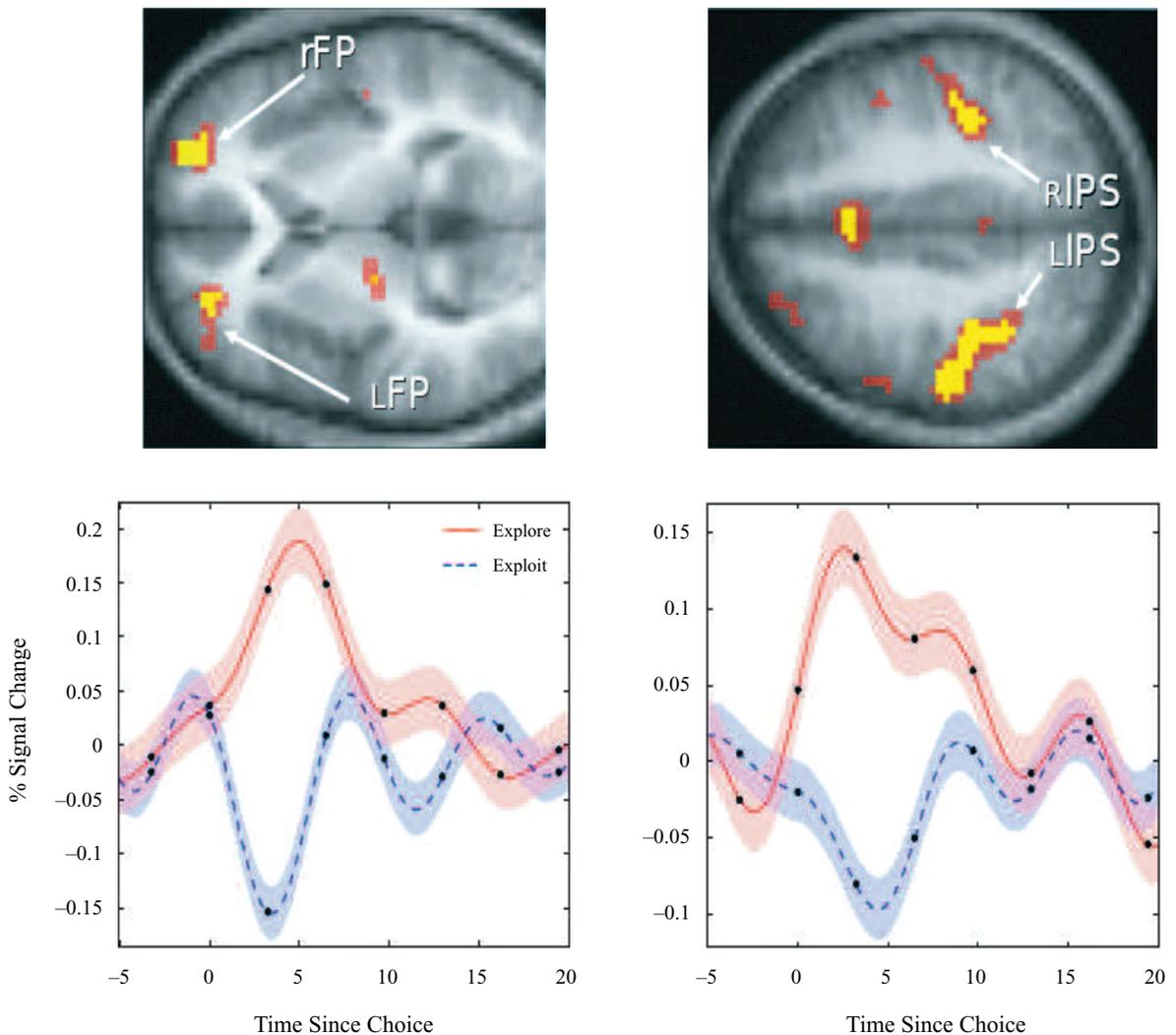
Figure 12. Areas differentially active during exploration. Top: Both the bilateral frontopolar cortex (rFP, lFP) and the anterior intraparietal sulcus (rIPS, lIPS) exhibited higher BOLD activity during choices estimated to be exploratory, as compared with exploitative ones. Bottom: BOLD time courses from these regions averaged over exploratory (solid) and exploitative (dashed) trials; both areas exhibit positive BOLD excursions during exploration and the opposite for exploitation. Adapted from "Cortical Substrates for Exploratory Decisions in Humans," by N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan, 2006, *Nature*, *441*, p. 878. Copyright 2006 by the Nature Publishing Group. Reprinted with permission.

of an option other than the one for which the subject was estimated to expect the most reward. Together with the behavioral results and other imaging findings suggesting the involvement of anterior prefrontal areas in processing uncertainty (e.g., Yoshida & Ishii, 2006) and of the intraparietal cortex in belief states (as discussed above), this neural dissociation suggests that rather than being encouraged by a bonus reported in a common currency with exploitative value, exploration in this task somehow draws on distinct neural processes.

It is certainly possible that high-level regulation of the sort associated with the frontal pole overrides a prepotent drive to exploit; however, we should stress that it is not clear from these data what function, if any, these areas causally contribute to exploration. It is also unclear how to reconcile these findings with the observation that dop-

aminergic neurons are activated by novelty (although, of course, BOLD activations are far from direct measurements of dopaminergic activity). One possibility, suggested by the Kakade and Dayan (2002) model and also by a follow-up imaging study (Wittmann, Daw, Seymour, & Dolan, 2008), is that the brain approximates the value of exploration by assigning bonuses for novelty, rather than for uncertainty. Novelty bonuses are particularly easy to compute—they require only optimistic initialization—but they are only an imperfect proxy for uncertainty. For instance, such bonuses would likely not be engaged by the task shown in Figure 1C and Figure 12, because the exploration there was mandated by changing reward probabilities in familiar bandits, rather than by explicit novelty.

Finally, if organisms indeed use some sort of random exploration, rather than exploration guided toward uncer-

tainty, the dynamic regulation of the degree of randomness becomes particularly crucial. Although there is no direct evidence how this might be conducted, theoretical speculation has focused on the neuromodulator norepinephrine as a potential substrate for such control (Cohen, McClure, & Yu, 2007; Doya, 2002; McClure, Gilzenrat, & Cohen, 2006).

CONCLUSIONS

We have reviewed some basic results in decision theory as they pertain to data in the psychology and neuroscience of choice. We considered the central computational issues surrounding the depths of the subjects' ignorance about the rules of the tasks they face and their immediate state within the task, and also the heights of their ambition as to whether to try to optimize exploration and exploitation jointly. We also considered a number of different algorithmic dimensions—most importantly, separating model-based algorithms—which make direct, computationally expensive use of the (perhaps estimated) rules of the task to work out optimal actions—and value-or policy-based model-free algorithms that do away with these complexities, at the expense of being less statistically efficient at turning information from the world into good actions. We illustrated these issues with a number of paradigmatic special cases in which we could also report relevant psychological and/or neural data.

A main focus of this review has been to highlight the commonalities among a large class of problems through the medium of Bayesian decision theory. Even though certain particular problems, such as the SPRT, admit particularly simple solutions (which can be analyzed by special methods; Shadlen et al., 2007), apparently straightforward extensions take us back to the general case. Broader solutions involve turning observations into beliefs about the state of the subject in the environment and handling sequential decision problems that involve optimization over multiple steps. We will end by considering some of the classes of question under current investigation.

From a computational perspective, the most interesting and pressing direction for future studies concerns the construction of a relevant state space. We have stressed the notion of a belief state in the context of a probabilistic model of a domain. This offers a critical, crisp foundation; however, it also poses a brace of challenging problems: a statistical one of determining such models from experience, and a computational one of doing inference in the face of the lurking intractability to which we have often referred. Even approximate inference is extremely hard. Defining, and finding, minimal models is an obvious direction; and indeed, there is much current interest in taking advantage of structural properties of ecologically relevant domains, such as various forms of hierarchy, in order to make progress with these problems. There is also a wealth of work in unsupervised learning that is a fertile source of ideas.

A second key computational and algorithmic topic is that of approximations in general. There are some obvious routes to creating approximately optimal policies—

for instance, restricting the number of basis functions in the representation of value functions, discretizing the belief state very coarsely, or restricting the length of the history of past observations used to create an effective state space. However, there are few broad results about the consequences of these approximations for the quality of control.

From a neural and psychological perspective, there are many open issues. One question under active debate involves the way in which uncertain information, such as belief states, might be represented in the activity of populations of neurons and support the basic computations of Bayesian decision theory, such as belief updating (Equation 26) and the calculation of expected values (Equation 22) (Beck, Ma, Latham, & Pouget, 2007; Beck & Pouget, 2007; Deneve, 2008; Jazayeri & Movshon, 2006; Ma et al., 2006; Rao, 2004; Sahani & Dayan, 2003; Zemel, Dayan, & Pouget, 1998). Conversely, it is as yet unclear to what extent subjects can perform these operations in rich domains (for instance, with multimodal posterior distributions), let alone get near to an optimal balance between exploration and exploitation.

A second important point concerns the existence and interaction of different classes of mechanisms and systems involved in decision making. We have mentioned studies in rats suggesting that model-based mechanisms and goal-directed control (involving the medial prefrontal cortex and the dorsomedial striatum) and model-free mechanisms and habitual control (involving the dorsolateral striatum and dopaminergic neuromodulation) coexist and, indeed, compete to control choices (Balleine et al., 2007; Daw et al., 2005; Dickinson & Balleine, 2002; Killcross & Coutureau, 2003). As far as exploitation is concerned, these two mechanisms represent two ends of a spectrum trading off the statistical efficiency of learning (favoring model-based control over model-free control, which learns by bootstrapping) for the computational efficiency of use (favoring model-free methods, which do not have to solve dynamic programming problems online). Other structures may also be involved—for instance, with the hippocampus contributing to control based on episodic memory (Lengyel & Dayan, 2008). The existence of discrete areas in exploration (Daw, O'Doherty, et al., 2006) is less expected from the perspective of Bayesian decision theory, in which the benefits of exploration apparent in Figure 9 are calculated of a piece with the benefits of exploitation.

To summarize, decision theory is one of the few areas in which there is a tight and productive coupling between (1) normative theory from statistics, operations research, artificial intelligence, economics, and engineering, (2) behavioral results in ethological and psychological paradigms, and (3) electrophysiological, pharmacological, and even anatomical neural data. In a surprising set of cases, algorithmic ideas from the former disciplines have found relatively direct psychological and neural instantiations in the latter ones, although this obviously need not be the case, particularly as the computations and algorithms get more complicated. The fruits of 50 years of analytical

research into decision making are being actively plucked in the form of biologically based models. Indeed, the high-quality choices made by animals and humans in environments replete with extreme computational challenges coming from uncertainty about states and rules are poised to provide a whole new impetus toward the theory of appropriate approximation.

REFERENCES

Ainslie, G. (2001). *Breakdown of will*. Cambridge: Cambridge University Press.

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *Journal of Neuroscience*, **27**, 8161-8165.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215-232). Cambridge, MA: MIT Press.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, **20**, 1391-1397.

Baxter, J., & Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, **15**, 319-350.

Beck, J. [M.], Ma, W. J., Latham, P. E., & Pouget, A. (2007). Probabilistic population codes and the exponential family of distributions. *Progress in Brain Research*, **165**, 509-519.

Beck, J. M., & Pouget, A. (2007). Exact inferences in a neural implementation of a hidden Markov model. *Neural Computation*, **19**, 1344-1361.

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology*, **191**, 391-431.

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments* (Monographs on Statistics and Applied Probability). London: Chapman & Hall.

Bertsekas, D. P. (2007). *Dynamic programming and optimal control* (2 vol.). Belmont, MA: Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychology Review*, **113**, 700-765.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience*, **13**, 87-100.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, **12**, 4745-4765.

Chrisman, L. (1992). Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the 10th National Conference on Artificial Intelligence* (pp. 183-188). Menlo Park, CA: AAAI Press.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B*, **362**, 933-942.

Costa, R. M. (2007). Plastic corticostriatal circuits for action learning: What's dopamine got to do with it? In B. W. Balleine, K. Doya, J. O'Doherty, & M. Sakagami (Eds.), *Reward and decision making in corticobasal ganglia networks* (Annals of the New York Academy of Sciences, Vol. 1104, pp. 172-191). New York: New York Academy of Sciences.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, **18**, 1637-1677.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, **16**, 199-204.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, **8**, 1704-1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, **441**, 876-879.

Day, J. J., Roitman, M. F., Wightman, R. M., & Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, **10**, 1020-1028.

Dayan, P., & Abbott, L. F. (2005). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dayan, P., & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, **25**, 5-22.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence* (pp. 761-768). Menlo Park, CA: AAAI Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1-38.

Deneve, S. (2008). Bayesian spiking neurons I: Inference. *Neural Computation*, **20**, 91-117.

Dickinson, A., & Balleine, B. (2002). The role of learning in motivation. In C. Gallistel (Ed.), *Stevens's handbook of experimental psychology* (Vol. 3, pp. 497-533). New York: Wiley.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, **15**, 495-506.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, **415**, 429-433.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, **8**, 1481-1489.

Friston, K. J., Tononi, G., Reeke, G. N., Jr., Sporns, O., & Edelman, G. M. (1994). Value-dependent selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, **59**, 229-243.

Gittins, J. C. (1989). *Multi-armed bandit allocation indices*. New York: Wiley.

Glimcher, P. W. (2004). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge, MA: MIT Press, Bradford Books.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, **5**, 10-16.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, **36**, 299-308.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, **30**, 535-574.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., et al. (2004). A neural correlate of reward-based behavioral learning in caudate nucleus: A functional magnetic resonance imag-

ing study of a stochastic decision task. *Journal of Neuroscience*, **24**, 1660-1665.

HYMAN, S. E., MALENKA, R. C., & NESTLER, E. J. (2006). Neural mechanisms of addiction: The role of reward-related learning and memory. *Annual Review of Neuroscience*, **29**, 565-598.

JAAKKOLA, T., JORDAN, M. I., & SINGH, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, **6**, 1185-1201.

JACOBS, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, **39**, 3621-3629.

JAZAYERI, M., & MOVSHON, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, **9**, 690-696.

JOEL, D., NIV, Y., & RUPPIN, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, **15**, 535-547.

KABLE, J. W., & GLIMCHER, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, **10**, 1625-1633.

KAELBLING, L. P. (1993). *Learning in embedded systems*. Cambridge, MA: MIT Press.

KAKADE, S., & DAYAN, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, **15**, 549-559.

KILLCROSS, S., & COUTUREAU, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, **13**, 400-408.

KÖRDING, K. [P.] (2007). Decision theory: What "should" the nervous system do? *Science*, **318**, 606-610.

KÖRDING, K. P., & WOLPERT, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, **427**, 244-247.

KREBS, J., KACELNIK, A., & TAYLOR, P. (1978). Test of optimal sampling by foraging great tits. *Nature*, **275**, 27-31.

LENGYEL, M., & DAYAN, P. (2008). Hippocampal contributions to control: The third way. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 889-896). Cambridge, MA: MIT Press.

LO, C.-C., & WANG, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, **9**, 956-963.

MA, W. J., BECK, J. M., LATHAM, P. E., & POUGET, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, **9**, 1432-1438.

MANGEL, M., & CLARK, C. W. (1989). *Dynamic modeling in behavioral ecology*. Princeton, NJ: Princeton University Press.

MCCLURE, S. M., GILZENRAT, M. S., & COHEN, J. D. (2006). An exploration–exploitation model based on norepinephrine and dopamine activity. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 867-874). Cambridge, MA: MIT Press.

MCCLURE, S. M., LAIBSON, D. I., LOEWENSTEIN, G., & COHEN, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, **306**, 503-507.

MCNAMARA, J., & HOUSTON, A. (1980). The application of statistical decision theory to animal behaviour. *Journal of Theoretical Biology*, **85**, 673-690.

MONTAGUE, [P.] R. (2006). *Why choose this book?: How we make decisions*. New York: Dutton.

MONTAGUE, P. R., DAYAN, P., & SEJNOWSKI, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, **16**, 1936-1947.

NEAL, R. M., & HINTON, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 355-368). Norwell, MA: Kluwer.

NG, A., HARADA, D., & RUSSELL, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 278-287). San Francisco: Morgan Kaufmann.

O'DOHERTY, J. P., DAYAN, P., FRISTON, K., CRITCHLEY, H., & DOLAN, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, **38**, 329-337.

O'DOHERTY, J. [P.], DAYAN, P., SCHULTZ, J., DEICHMANN, R., FRIS-

ton, K., & DOLAN, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**, 452-454.

PARKER, A. J., & NEWSOME, W. T. (1998). Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience*, **21**, 227-277.

PLATT, M. L., & GLIMCHER, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, **400**, 233-238.

PUTERMAN, M. L. (2005). *Markov decision processes: Discrete stochastic dynamic programming*. New York: Wiley Interscience.

PYKE, G. H. (1984). Optimal foraging theory: A critical review. *Annual Review of Ecology & Systematics*, **15**, 523-575.

RAO, R. P. N. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, **16**, 1-38.

RATCLIFF, R., & McKOON, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, **20**, 873-922.

RATCLIFF, R., & ROUDER, J. (1998). Modeling response times for two-choice decisions. *Psychological Science*, **9**, 347-356.

REDGRAVE, P., GURNEY, K., & REYNOLDS, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews*, **58**, 322-339.

RESCORLA, R., & WAGNER, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

ROITMAN, J. D., & SHADLEN, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, **22**, 9475-9489.

ROSS, S. (1983). *Introduction to stochastic dynamic programming: Probability and mathematical*. Orlando, FL: Academic Press.

SAHANI, M., & DAYAN, P. (2003). Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Computation*, **15**, 2255-2279.

SCHULTZ, W. (2002). Getting formal with dopamine and reward. *Neuron*, **36**, 241-263.

SCHULTZ, W., DAYAN, P., & MONTAGUE, P. R. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593-1599.

SEYMOUR, B., O'DOHERTY, J. P., DAYAN, P., KOLTZENBURG, M., JONES, A. K., DOLAN, R. J., ET AL. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, **429**, 664-667.

SHADLEN, M. N., BRITTEN, K. H., NEWSOME, W. T., & MOVSHON, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, **16**, 1486-1510.

SHADLEN, M. N., HANKS, T. D., CHURCHLAND, A. K., KIANI, R., & YANG, T. (2007). The speed and accuracy of a simple perceptual decision: A mathematical primer. In K. Doya, S. Ishii, A. Pouget, & R. P. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding* (pp. 209-238). Cambridge, MA: MIT Press.

SHADLEN, M. N., & NEWSOME, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences*, **93**, 628-633.

SMITH, P. L., & RATCLIFF, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, **27**, 161-168.

STOCKER, A. A., & SIMONCELLI, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, **9**, 578-585.

SURI, R. E., & SCHULTZ, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, **121**, 350-354.

SUTTON, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3**, 9-44.

SUTTON, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In B. W. Porter & R. J. Mooney (Eds.), *Proceedings of the Seventh International Conference on Machine Learning* (pp. 216-224). San Francisco: Morgan Kaufmann.

SUTTON, R. S., & BARTO, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

TROMMERSHÄUSER, J., LANDY, M. S., & MALONEY, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychological Science*, **17**, 981-988.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003a). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America A*, **20**, 1419-1433.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003b). Statistical decision theory and trade-offs in the control of motor response. *Spatial Vision*, **16**, 255-275.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, **108**, 550-592.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, **36**, 955-968.

Wang, X.-J. (2006). Toward a prefrontal microcircuit model for cognitive deficits in schizophrenia. *Pharmacopsychiatry*, **39**(Suppl. 1), S80-S87.

Watkins, C. (1989). *Learning from delayed rewards*. Unpublished doctoral thesis, University of Cambridge.

Whiteley, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision*, **8**, 1-15.

Wickens, J. [R.] (1990). Striatal dopamine in motor activation and reward-mediated learning: Steps towards a unifying model. *Journal of Neural Transmission*, **80**, 9-31.

Wickens, J. R., Horvitz, J. C., Costa, R. M., & Killcross, S. (2007). Dopaminergic mechanisms in actions and habits. *Journal of Neuroscience*, **27**, 8181-8183.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8**, 229-256.

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, **58**, 967-973.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, **447**, 1075-1080.

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*, **50**, 781-789.

Yu, A. J. (2007). Optimal change-detection and spiking neurons. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems 19* (pp. 1545-1552). Cambridge, MA: MIT Press.

Yuille, A. J., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123-161). Cambridge: Cambridge University Press.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, **10**, 403-430.