Decision Theory, Reinforcement Learning and the Brain

Peter Dayan
Gatsby Computational Neuroscience Unit
UCL
dayan@gatsby.ucl.ac.uk

Nathaniel D. Daw
Neural Science & Psychology
NYU
nathaniel.daw@nyu.edu

Abstract

Decision theory 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; algorithmic topics addressing model-based and model-free methods for making choices; and highlight key aspects of the neural implementation of decision-making.

1 Introduction

The abilities of animals to make predictions about the affective nature of their environments and to exert control 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 (*eg* Bellman, 1957; Puterman, 2005; Berger, 1985; Bertsekas and Tsitsiklis, 1996; Bertsekas, 2007; Sutton and Barto, 1998; Green and Swets, 1966; Mangel and Clark, 1989; Montague, 2006; Gold and Shadlen, 2002, 2007; Glimcher, 2004; Körding, 2007; Gittins, 1989; Berry and Fristedt, 1985; Wald, 1947; Yuille and Bülthoff, 1996; McNamara and Houston, 1980), and is therefore a critical tool for modeling, understanding and predicting psychological data and its 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 were wired up to a device that delivered variable strength electric shocks. The delivery of the shocks was preceded by visual cues (cue A through cue D) in a sequence. Cue A occurred on 50% of trials; it was followed by cue B and then a larger shock 80% of the time; and by cue D and then a smaller shock 20% of the time. The converse was 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 section 3.1; 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 and Barto, 1998).

[Figure 1 about here.]

Figure 1B depicts a decision task that is closely related to signal detection theory (Green and Swets, 1966), and has been particularly illuminating about the link between neural activity and perception (Britten et al., 1992, 1996; Shadlen et al., 1996; Shadlen and Newsome, 1996; Gold and Shadlen, 2001, 2002, 2007). In the classical version of this task, monkeys watch a screen that shows moving dots. Some proportion of the dots are moving in one direction; the rest are moving in 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 sections 3.2 and 3.3, along with the roles of two visual cortical areas (MT and 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 Markov decision problem (POMDP) related to the sequential probability ratio test (SPRT; Wald, 1947; Smith and Ratcliff, 2004; Ratcliff and Rouder, 1998; Shadlen et al., 2007).

Finally, figure 1C shows a further decision-theoretic wrinkle in the form of a experiment into the tradeoff between exploration and exploitation Daw et al. (2006b). 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 only find out about the current value of a machine by choosing it; and so have to balance picking the machine which 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 (Gittins, 1989; Berry and Fristedt, 1985); section 3.4 discusses the issues and approximate solutions, including one that, evidence suggests, implicates 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 which 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 are thereby integrated, and common solution mechanisms identified. In particular, viewing these problems as different specializations of a common task involving both sensory inference and learning 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 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 consideration of figure 2 concern what the subjects *know* and what they are trying to *accomplish*. The prediction task of 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 of figure 1B originate in a case that subjects know the rules of the task but are only incompletely certain about the state. The exploration/exploitation tradeoff of figure 1C can be seen as combining both of these in a case that subjects are ambitious about behaving optimally in the face of whatever uncertainty they have. Critically, through the medium of the task 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 that information from past and present trials can be accumulated; and thence to implementational issues in terms of the neural structures involved in the solutions.

[Figure 2 about here.]

2 Foundational Issues

The task in figure 2 only involves four choice points or *states* (x_1, x_2, x_3, x_4) signalled, perhaps imperfectly (*ie* leaving some uncertainty, in a way we will formulate precisely later) by cues (c_1, c_2, c_3, 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, and with C being costly. The subjects may not know these payoffs at the outset.

Computational Issues

As 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 (*ie* which x_i they currently occupy), and/or the rules of the task (*ie* 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) then 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 also more persistently 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, *ie* trying to make the reward for the current trial as valuable as possible, given whatever they currently know about the task. In the case that 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 (*eg* Ainslie, 2001; McClure et al., 2004; Kable and Glimcher, 2007).

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 that it could also 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, *ie* have algorithms that only require 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 non-neurobiological systems.

Algorithms differ in how they draw on experience to estimate quantities relevant to the decision, and how they render these into choices. The most important algorithmic dimension is that distinguishing *model-based* and *model-free* methods (Sutton and 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 is clearly enough just to know four letters, *viz* 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 derived from other information, like the expected future utilities ("values") that will accrue from different actions or states, or

(in model-based methods) from the rules of the task. Indeed, one of the most important products of the field of reinforcement learning (Sutton and 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 based on 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 (Wickens, 1990; Friston et al., 1994; Barto, 1995; Montague et al., 1996; Schultz et al., 1997; Suri and Schultz, 1998; Joel et al., 2002; Daw et al., 2005). For aversive outcomes such as the shocks of 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 prefrontal cortex and dorsomedial striatum (Dickinson and Balleine, 2002; Everitt and Robbins, 2005; Balleine et al., 2007; Daw et al., 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 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 ex-

tensively investigated in a rather different set of psychophysical tasks (Britten et al., 1992, 1996; Parker and Newsome, 1998; Platt and Glimcher, 1999; Gold and Shadlen, 2007), 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 cortex. However, this work has generally confined itself to fairly rudimentary and limited forms of learning.

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

3 Examples

We use the examples of figure 1 rendered in the abstract forms of 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 and Newsome, 1996), information gathering and optimal stopping (Gold and Shadlen, 2001, 2007) and the exploration/exploitation tradeoff (Daw et al., 2006b). 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 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.

3.1 The Markov decision problem

The central problem in the prediction task of figure 1A is that until the subject observes cue A or C at the start of a trial, she does not know whether 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 the 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 of figure 3A in which the cues unambiguously identify the state (c_1 for x_1, and so on). We 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 $a \in \{\mathrm{L}, \mathrm{R}, \mathrm{C}\}$ to each state x_i. The probabilities $p_{ij} = p_{ij}(\mathrm{C})$ indicate the probabilities of going from state x_i to x_j under action C, whose cost is $r_i(\mathrm{C})$; actions L and R have deterministic consequences. Exactly how the 'best' policy is defined depends upon the particular goal. For now, we will assume that the rewards and costs earned across each whole single trial are simply summed, and this sum is what has to be predicted and optimized.

[Figure 3 about here.]

First, consider the case that 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 the best action, *ie* one that maximises this expected return is

$$\pi_1^* = \mathrm{argmax}_{a\in\{\mathrm{L},\mathrm{R}\}}\,[Q_1^*(a)] = \mathrm{L} \quad \text{and similarly} \quad \pi_2^* = \mathrm{argmax}_{a\in\{\mathrm{L},\mathrm{R}\}}\,[Q_2^*(a)] = \mathrm{R} \tag{2}$$

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, then 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; Ross, 1983; Bertsekas, 2007) fashions this observation into computational and algorithmic methods, which themselves underly reinforcement learning (RL). 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

$$V_1^* = \max_{a\in\{\mathrm{L,R}\}}\left[Q_1^*(a)\right] = 2 \quad \text{and similarly at } x_2 \quad V_2^* = \max_{a\in\{\mathrm{L,R}\}}\left[Q_2^*(a)\right] = 2 \tag{3}$$

which are shown in figure 3B. These values will be available provided the subjects choose correctly (*ie* 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

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

since the reward for the first action (at state x_3; $r_3(\mathrm{C})$) 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, which multiply the values V_1^* and V_2^* indicate the reward that can be achieved starting from those states, given appropriate actions there.

The values of L and R at x_3 are just $Q_3^*(\mathrm{L}) = r_3(\mathrm{L})$ and $Q_3^*(\mathrm{R}) = r_3(\mathrm{R})$, and so the optimal policy at x_3 is, as in equation 2,

$$\pi_3^* = \mathrm{argmax}_{a\in\{\mathrm{L,R,C}\}}\left[Q_3^*(a)\right] = \mathrm{C} \quad \text{and similarly} \quad \pi_4^* = \mathrm{argmax}_{a\in\{\mathrm{L,R,C}\}}\left[Q_4^*(a)\right] = \mathrm{C} \tag{5}$$

That is, in the example in figure 3A-C, 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 only showed the computations underlying the simplest dynamic programming problem, solving more realistic cases which nevertheless retain the structure of our task is a straightforward extension (see, for example, Bertsekas, 2007; Sutton and 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 only depend on the current state, and not 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 are not determinate of the state. This issue is central to the remaining examples in the paper.

We have so far assumed that the subjects know the rules of the task. In section 3.4 we consider the (much more involved) computational problem that arises in the case that 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 ($r_3(\mathrm{C})$, 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 (Puterman, 2005; Bertsekas, 2007; Sutton and Barto, 1998). The key observation is that equation 4, which is one of a number of forms of 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 equation 3. 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 that the subject performs action C at state x_3 a total of M times, getting to states $x_{j^1} \dots x_{j^M}$, for $j^1, \dots, j^M \in \{1, 2\}$ sampled from $p_{31}; p_{32}$, and experiencing rewards $r^1 \dots r^M$, sampled from $r_3(\mathrm{C})$. Then, one might estimate $r_3(\mathrm{C})$ and p_{31} by the sample mean estimates

$$\hat{r}_3(\mathrm{C}) = \frac{1}{M}\sum_{k=1}^{M} r^k \quad \text{and} \quad \hat{p}_{31} = \frac{1}{M}\sum_{k=1}^{M} \chi\left(x_{j^k}, x_1\right) \ , \tag{6}$$

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 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 (RL) 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 differ-

ence 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

$$\delta^k = r_3(\mathrm{C}) + V^*_{j^k} - Q^*_3(\mathrm{C}) \tag{7}$$

for trial k in which the transition went from state x_3 to state x_{j^k}. Such errors can be used to update value estimates (in this case of $Q^*_3(\mathrm{C})$) to reduce inconsistencies. These algorithms are guaranteed to converge to optimal value functions (which determine optimal policies) in the limit of indefinite sampling (Watkins, 1989; Jaakkola et al., 1994; Bertsekas and Tsitsiklis, 1996). However, given only finite experience, the value estimates will again be subject to sampling noise, and so 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 $Q^*_3(\mathrm{C})$) based on other estimates ($V^*_{j^k}$). This makes them statistically inefficient, since early on, estimates such as $V^*_{j^k}$ are inaccurate themselves, and so can only support poor learning.

Although these algorithms are model-free, (*ie* 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 states in terms of the summed reward that is expected to accrue across the whole rest of the trial. Apart from these value-based RL 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 (Williams, 1992; Baxter and Bartlett, 2001) 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

$$P_1(a = \mathrm{L}; w_1) = \pi_1(w_1) = \sigma(w_1) \quad \text{where} \quad \sigma(w) = 1/(1 + \exp(-w)) \tag{8}$$

is the standard logistic sigmoid function. Naturally, $P_1(a = \mathrm{R}; w_1) = 1 - \pi_1(w_1) = \sigma(-w_1)$. 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

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

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

$$w_1 \Rightarrow w_1 + \epsilon \Delta w_1^k \quad \text{where} \quad \Delta w_1^k = r^k (1 - P_1(a = a^k; w_1)) \tag{9}$$

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

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

and so the latter quantity should increase in the average, at least provided 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 favoured. As advertized, 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 based on the reward $r_3(\mathrm{L})$ that is only available at a later point.

Psychology & Neuroscience

There is a wealth of work in both psychology and neuroscience on tasks that can be considered as 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 concerned predicting reinforcement in tasks without decisions.

Perhaps the best developed connection between these ideas and neural data is through predic-

tion errors for model-free RL, as in equation 7. Versions 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.

[Figure 4 about here.]

Consider the experiment of figure 1A. The task was designed to induce higher-order prediction errors, *ie*, those arising from changes in expectations about future reinforcement rather than 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 timeseries 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 though to be likely will not occur; this is negative prediction error (and a relative decrease in BOLD; figure 4C). Figure 4D illustrates how we can extend the same logic a further step back, just as in the dynamic programming analysis of Markov decision processes. Here, since cue A indicates that cue B (and thence the large shock) is likely, it also induces positive prediction error when it appears, signaling an end to the relatively safe period between trials.

Seymour et al. (2004)'s study is in the aversive domain. For appetitive outcomes, there is ample evidence that the phasic activity of dopamine cells in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) in monkeys (*eg* Schultz, 2002), and the release of dopamine at

striatal targets in rats (Day et al., 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 behaviour (see, for some recent highlights, Joel et al., 2002; Hyman et al., 2006; Wickens et al., 2007; Costa, 2007). The proposal that this operates according to the rules of reinforcement learning (O'Doherty et al., 2003, 2004; Barto, 1995; Montague et al., 1996; Schultz et al., 1997; Suri and Schultz, 1998; Joel et al., 2002; Balleine et al., 2007; Daw and Doya, 2006; Haruno et al., 2004) in a way that ties together the at least equally extensive data on the psychology of instrumental choice with these neural data, has extensive, though not universal, support (*eg* Berridge, 2007; Redgrave et al., 2007).

However, behaviorally sophisticated experiments (reviewed in Dickinson and Balleine, 2002; Balleine et al., 2007) 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 only change 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; and model-free choices (called *habits*) evident after more extensive experience, simpler tasks, and actions further from outcomes. Furthermore, these two forms of control can be differentially suppressed by selective lesions of parts of medial prefrontal cortex in rats (Killcross and Coutureau, 2003). Daw et al. (2005) argued that the tradeoff 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.

Though more comprehensive, even this synthesis has an extremely limited scope. As 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 turn to these.

3.2 Signal detection theory

The task shown in figure 1B, in the version that 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 that the rules are known, but the inputs c_i 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 and 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 *hidden state,* since the subject occupies a true state in the world which 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.

[Figure 5 about here.]

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:

$$p_1(c_\alpha = c | x_\alpha = x_1) \equiv p_1(c_\alpha | x_1) \qquad \text{and} \qquad p_2(c_\alpha = c | x_\alpha = x_2) \equiv p_2(c_\alpha | x_2)$$

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 figure 5A;B.

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 the table:

	action	
actual state	L	R
x_1	**hit**	**miss**
x_2	**false alarm**	**correct rejection**

although 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.

[Figure 6 about here.]

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, *ie* the posterior distribution over being in x_1 or x_2 given the data:

$$P(x_\alpha = x_1|c_\alpha) \equiv P(x_1|c_\alpha) = \frac{p_1(c_\alpha|x_1)P(x_1)}{p(c_\alpha)} = \frac{1}{1 + \frac{1}{l(c_\alpha)}\frac{P(x_2)}{P(x_1)}} \tag{10}$$

where

$$l(c_\alpha) = \frac{p_1(c_\alpha|x_1)}{p_2(c_\alpha|x_2)} \tag{11}$$

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 that 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 $P(x_1) = 0.5$). Errors occur where the posterior is uncertain, near $P(x_1|c_\alpha) = 0.5$. 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):

$$Q^*_{c_\alpha}(\mathrm{L}) = \mathcal{E}[r_{c_\alpha}(\mathrm{L})|c_\alpha] = P(x_1|c_\alpha)r_1(\mathrm{L}) + P(x_2|c_\alpha)r_2(\mathrm{L}) \tag{12}$$

$$Q^*_{c_\alpha}(\mathrm{R}) = \mathcal{E}[r_{c_\alpha}(\mathrm{R})|c_\alpha] = P(x_1|c_\alpha)r_1(\mathrm{R}) + P(x_2|c_\alpha)r_2(\mathrm{R}) \tag{13}$$

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 values, then, as in equation 2, we can choose an optimal policy

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

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 be written as

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

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

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

$$\theta_B = \frac{r_2(\text{R}) - r_2(\text{L})}{r_1(\text{L}) - r_1(\text{R})} \frac{P(x_2)}{P(x_1)} \tag{16}$$

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

To summarize, Bayesian decision theory in the case of state uncertainty is formally just the same as in the case of complete knowledge of the state, except redefining the state to be the belief state $P(x_\alpha = x_1|c_\alpha)$. Given this, the same ideas as 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 and 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, *ie* 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 a discrete quantity like x_1 in the fully observable MDP. Further, although the 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 that $\sigma_1 = \sigma_2$, the optimal decision can also be described by a threshold in the observable quantity c_α. However, if $\sigma_1 \neq \sigma_2$, then, 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

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

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

$$Q^*_{c_\alpha}(\mathrm{L}) = \sum_k f_k(c_\alpha) w_k(\mathrm{L}) \tag{18}$$

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 & 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 and Simoncelli, 2006), and changes in the observations (Körding and Wolpert, 2004); subjects even appear to have a good idea about the noise associated with their own sensations (Whiteley and Sahani, 2008) and actions (Trommershäuser et al., 2003b,a, 2006), and can cope with even more sophisticated cases in which cues are two dimensional (visual: c^v_α and auditory: c^a_α) and are conditionally independent given the state (Ernst and Banks, 2002; Battaglia et al., 2003; Jacobs, 1999; Yuille and 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. (1992, 1996); 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). Some proportion of the dots all moves coherently, in the same direction, the remainder move at random, and the job of the monkey is to report which direction the coherently moving dots favour based on 2 seconds' worth of observing the motion. Typically, the monkeys only have two choices, 180° apart, *ie* 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 7 about here.]

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 (*ie* the percentage of the random dots moving in a coherent 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 these 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 shows 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 are made by the whole monkey. Of course, the monkey's problem is to pick out the cells of this calibre (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 should mitigate our surprise that the overall performance of the monkey is not substantially better than that of a single, somewhat randomly recorded, neuron.

3.3 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 partially observable Markov decision problems (POMDPs). This is exemplified by the other version of the task in figure 1B, in which the 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

Figure 8A;B show a version of the task which combines some of the Markov decision problem aspects of section 3.1 with the state uncertainty of section 3.2. Here, the 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. Subjects could either choose L (which is correct at x_3), R (correct at x_4), or they could choose C, which incurs a small penalty (-0.1), but delivers them to $x_\alpha \in \{x_1, x_2\}$. It might be wise, if the cue available there c_α (assumed to be suitably independent of c_β) better resolves their state uncertainty (*ie* 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 as a form of exploration.

[Figure 8 about here.]

The Bayesian decision-theoretic ideas articulated in section 3.1 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 consider the evolution of the belief state, and then 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 as for x_1 and x_2 in section 3.2. 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 if the subject chooses C at the first stage, then it will observe c_α at $x_\alpha \in \{x_1, x_2\}$; given this observation, it is again 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:

$$\begin{aligned} P(x_1|c_\beta, c_\alpha) &\propto p_1(c_\alpha|x_1)P(x_3|c_\beta) \\ &= p_1(c_\alpha|x_1)p_3(c_\beta|x_3)p(x_3) \end{aligned} \tag{19}$$

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 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.

[Figure 9 about here.]

To calculate the optimal policy, we now proceed exactly as in section 3.1, *backwards* from the second state to the first. If the subject gets to x_1 or x_2, it will be able to make its choice based on its 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

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

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 $x_\alpha = x_2$, *ie* when it can be sure which action to perform. If instead the cues are inconsistent, then the value is closer to 0.5, 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, she 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 favours x_3, then 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:

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

$$= P(x_3|c_\beta)\int_{c_\alpha} V^*_{\alpha,c_\alpha,c_\beta} p(c_\alpha|x_1)dc_\alpha + P(x_4|c_\beta)\int_{c_\alpha} V^*_{\alpha,c_\alpha,c_\beta} p(c_\alpha|x_2)dc_\alpha \tag{22}$$

since the transitions can only occur from $x_3 \to x_1$ or $x_4 \to x_2$. This is the exact equivalent of equation 3, 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:

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

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 $1 + r_{c_\beta}(\mathrm{C}) = 0.9$, given the cost -0.1 of probing.

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

$$Q^*_{\beta,c_\beta}(\mathrm{L}) = P(x_3|c_\beta) \qquad Q^*_{\beta,c_\beta}(\mathrm{R}) = P(x_4|c_\beta) \tag{24}$$

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 above equations, the subject should choose C if

$$\max\{P(x_3|c_\beta), P(x_4|c_\beta)\} < -0.1 + \mathcal{E}_{p(c_\alpha|c_\beta)}\left[\max\{P(x_1|c_\alpha, c_\beta), P(x_2|c_\alpha, c_\beta)\}\right], \tag{25}$$

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 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 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, future) of 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 history 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 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 $P(x_\tau|\{c_1, \ldots c_\tau\})$, 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 used to learn values or policies on the basis of it (Chrisman, 1992; Daw et al., 2006a). 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 using belief states, 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 (Rao, 2004; Ma et al., 2006), 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 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, then equation 19 would sum over both possibilities:

$$P(x_1|c_\beta, c_\alpha) \propto p_1(c_\alpha|x_1)P(x_1 \mid x_3)P(x_3|c_\beta) + p_1(c_\alpha \mid x_1)P(x_1 \mid x_4)P(x_4 \mid c_\beta) \tag{26}$$

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 6); 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 et al., 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 equation 6. 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 and 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 & Neuroscience

Drift diffusion decision-making

In the original studies described in figure 1B and figure 7 that linked the activity of MT neurons to choice behavior, the monkeys observed the random dot motion displays for a fixed period of 2 seconds 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 sequential probability ratio test (SPRT Wald, 1947; Gold and Shadlen, 2001, 2007; Smith and Ratcliff, 2004; Ratcliff and Rouder, 1998; Shadlen et al., 2007), 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: $\ldots, c_\tau, \ldots$), pertaining to a binary discrimination (in our case between x_3, x_1, which both 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:

$$\ell_T(\mathbf{c}_T) = \log \frac{p(c_1, ..., c_T \mid x_\mathrm{L})}{p(c_1, ..., c_T \mid x_\mathrm{R})} = \sum_\tau \log \frac{p(c_\tau | x_L)}{p(c_\tau | x_R)} = \sum_\tau \ell_\tau(c_\tau)$$

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:

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

or equivalently 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 figure 8A;B 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 which will look like a threshold on the belief state.

Figure 10 shows schematic results of an experiment (cartooned in part 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 (Roitman and Shadlen, 2002; Gold and Shadlen, 2001, 2007). The main curves in figures 10B;C show 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 grey patches in figure 10A); that is they fire selectively when subjects are planning saccades to targets in space of the sort employed this experiment. The curves in figures 10B;C show the mean firing rates of such neurons under various conditions over the course of trials.

[Figure 10 about here.]

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 (colours) 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 (gold), following a (different) transient, this rises steeply. For less coherent motion (red), this rises less steeply; for incoherent motion (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 an log likelihood ratio computed by integrating opponent activity from MT neurons with opposing preferences. 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 favours this direction.

Figure 10C shows the same conditions, but now triggered on the saccades themselves. The curves all 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 in section 3.2 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, a wealth of very important studies has used it and related models (*eg* Smith and Ratcliff, 2004; Ratcliff and McKoon, 2008) 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 is also a number of influential connectionist (Usher and McClelland, 2001) and neural (Wang, 2002, 2006; Lo and Wang, 2006) models of this, together with mathematical theory about their interrelationship (Bogacz et al., 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 non-constant thresholds (*eg* Platt and Glimcher, 1999; Glimcher, 2004; 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 and 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 already 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 sug-

gests broadly how the brain might solve the policy learning problem for it (Gold and Shadlen, 2002, Ahmadi & Latham, personal communication): by using general reinforcement learning techniques — such as those described in section 3.1 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 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 et al., 2006a).

3.4 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 of figure 1C.

Doing so requires balancing exploitation against *exploration*: 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 we treated ignorance of the state — by planning on the basis of beliefs rather than directly on observations directly. In turn, the value of exploration can be quantified just as we have previously evaluated information-gathering ("probing") actions like C.

The computational issue

[Figure 11 about here.]

Consider the one-state (x_1) example of 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 its 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, then clearly the expected value of either is 0.5 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, then 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.

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 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:

$$E[p_L] = \frac{r_L + 1}{r_L + u_L + 2} \tag{28}$$

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 $Q_{\langle 0,0,0,0\rangle}(\mathrm{L})$ 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: i.e., 50%. That is, equation 28 does not just give the subject's *best guess* as to the rules: since it arises from 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 towards the same sort of recursion as 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

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

or more generally

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

and similarly for the value of R.

Finally, just as in equation 3, the value V^* of a belief state is just that of the better choice there:

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

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 tradeoff. 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, farther from the end of the game, exploration has value. To see why, consider a game in which p_R is known to be exactly 0.5 and we focus only on learning p_L. At the start of the game, the expected immediate payoff of 0.5 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 0.5, then 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 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 $V^*_{\langle 0,1,0,0\rangle}$ can't 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 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 which were generated by giving the agent 7 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 percent. 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, exact solution to even the smallest such problems is intractable, due to 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 RL algorithms discussed in 3.1 to work; but this can just as easily be viewed as a simplifying assumption that precludes optimal exploration, rather than 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 (though still non-trivial) subproblem. The approach takes advantage of a sort of independence between the actions arising from the problem structure. It works for multi-armed bandit problems of the sort described above, though only when the number of trials N is either infinite (and future reward is discounted

exponentially) or unknown with a constant hazard function, *ie* 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 (Sutton, 1990; Kaelbling, 1993; Dayan and Sejnowski, 1996; Dearden et al., 1998). 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 (*eg* Ng et al., 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 factors such 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.

Neuroscience and psychology

Outside the field of ethology, (*eg* McNamara and Houston, 1980; Pyke, 1984; Mangel and Clark, 1989) exemplified by some early work on the exploration of birds (Krebs et al., 1978), there is relatively little direct evidence as to how animals or humans explore. There is, 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 et al. (2006b) 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 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, 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

$$P_1(a = \mathrm{L}) = \sigma\left(\beta\left(Q_{\langle r_L,u_L,r_R,u_R\rangle}(L) - Q_{\langle r_L,u_L,r_R,u_R\rangle}(R)\right)\right)$$

where $\beta > 0$ regulates the strength of competition, *ie* 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.

[Figure 12 about here.]

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 like that of exploitation, for instance through dopaminergic spiking. Contrary to this expectation, Daw et al. (2006b) found that while dopaminergic efferent areas such as 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 intraparietal cortex — were activated preferentially when subjects explored rather than exploited (figure 12). Here, exploration was operationally defined as a choice 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 (*eg* Yoshida and Ishii, 2006) and of 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 dopaminergic neurons are activated by novelty (though, 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 followup imaging study (Wittmann et al., 2008), is that the brain approximates the value of exploration by assigning bonuses for novelty rather than uncertainty. Novelty bonuses are particularly easy to compute — they only require optimistic initialization — but they are only an imperfect proxy for uncertainty. For instance, such bonuses would likely not be engaged by the task of figure 1C and figure 12, because the exploration there was mandated by changing reward probabilities in familiar bandits rather than explicit novelty.

Finally, if organisms indeed use some sort of random exploration rather than exploration guided toward uncertainty, then the dynamic regulation of the degree of randomness becomes particularly crucial. While there is no direct evidence how this might be conducted, theoretical speculation has focused on the neuromodulator norepinephrine (NE) as a potential substrate for such control Doya (2002); McClure et al. (2006); Cohen et al. (2007).

4 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 and jointly optimize exploration and exploitation. 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 (that 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 its environment, and handling sequential decision problems which involve optimization over multiple steps. We 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 computaitonal 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, discretising 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 is the way that 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) (Zemel et al., 1998; Sahani and Dayan, 2003; Ma et al., 2006; Deneve, 2008; Rao, 2004; Jazayeri and Movshon, 2006; Beck et al., 2007; Beck and Pouget, 2007). Conversely, it is as yet unclear the extent to which subjects can perform these operations in rich domains (for instance with multimodal posterior distributions) let alone getting 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 dorsolateral 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 (Dickinson and Balleine, 2002; Balleine et al., 2007; Daw et al., 2005; Killcross and 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 and Dayan, 2008). The existence of discrete areas in exploration (Daw et al., 2006b) 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 normative theory from statistics, operations research, artificial intelligence, economics and engineering, behavioural results in ethological and psychological paradigms, and 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 case, particularly as the computations and algorithms get more complicated. The fruits of fifty years of analytical research into decision-making are being actively reaped in the form of biologically-based models – 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 towards the theory of appropriate approximation.

Acknowledgements

Funding came from the Gatsby Charitable Foundation (PD). We are grateful to our many collaborators and colleagues associated with our studies in these fields, notably Peter Latham, Read Montague, Yael Niv, Alex Pouget, Maneesh Sahani and Mike Shadlen. We also thank the editors and anonymous reviewers for their help comments.

References

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

Balleine, B. W., Delgado, M. R., and Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *J Neurosci*, 27(31):8161–8165.

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

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1391–1397.

Baxter, J. and Bartlett, P. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15(4):319–350.

Beck, J., Ma, W. J., Latham, P. E., and Pouget, A. (2007). Probabilistic population codes and the exponential family of distributions. *Prog Brain Res*, 165:509–519.

Beck, J. M. and Pouget, A. (2007). Exact inferences in a neural implementation of a hidden markov model. *Neural Comput*, 19(5):1344–1361.

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

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl)*, 191(3):391–431.

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments (Monographs on Statistics and Applied Probability) (Monographs on Statistics and Applied Probability)*. Springer.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control (2 Vol Set)*. Athena Scientific.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*. Athena Scientific.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev*, 113(4):700–765.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., and Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Vis Neurosci*, 13(1):87–100.

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

Chrisman, L. (1992). Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctions Approach. In *Proceedings of AAAI 10*, pages 183–188. Amer Assn for Artificial Intelligence.

Cohen, J. D., McClure, S. M., and Yu, A. J. (2007). Should i stay or should i go? how the human

brain manages the trade-off between exploitation and exploration. *Philos Trans R Soc Lond B Biol Sci*, 362(1481):933–942.

Costa, R. M. (2007). Plastic corticostriatal circuits for action learning: what's dopamine got to do with it? *Ann N Y Acad Sci*, 1104:172–191.

Daw, N. D., Courville, A. C., Tourtezky, D. S., and Touretzky, D. S. (2006a). Representation and timing in theories of the dopamine system. *Neural Comput*, 18(7):1637–1677.

Daw, N. D. and Doya, K. (2006). The computational neurobiology of learning and reward. *Curr Opin Neurobiol*, 16(2):199–204.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006b). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879.

Day, J. J., Roitman, M. F., Wightman, R. M., and Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat Neurosci*, 10(8):1020–1028.

Dayan, P. and Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA.

Dayan, P. and Sejnowski, T. (1996). Exploration Bonuses and Dual Control. *Machine Learning*, 25(1):5–22.

Dearden, R., Friedman, N., and Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence table of contents*, pages 761–768. American Association for Artificial Intelligence Menlo Park, CA, USA.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Deneve, S. (2008). Bayesian spiking neurons I: inference. *Neural Comput*, 20(1):91–117.

Dickinson, A. and Balleine, B. (2002). The role of learning in motivation. In Gallistel, C., editor, *Stevens' handbook of experimental psychology*, volume 3, pages 497–533, New York, NY. Wiley.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, 15(4-6):495–506.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statis-

tically optimal fashion. *Nature*, 415(6870):429–433.

Everitt, B. J. and Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat Neurosci*, 8(11):1481–1489.

Friston, K. J., Tononi, G., Reeke, G. N., Sporns, O., and Edelman, G. M. (1994). Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience*, 59(2):229–243.

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices (Wiley Interscience Series in Systems and Optimization)*. John Wiley & Sons Inc.

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

Gold, J. and Shadlen, M. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends Cogn Sci*, 5(1):10–16.

Gold, J. I. and Shadlen, M. N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2):299–308.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annu Rev Neurosci*, 30:535–574.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons.

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., Imamizu, H., and Kawato, M. (2004). A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J Neurosci*, 24(7):1660–1665.

Hyman, S. E., Malenka, R. C., and Nestler, E. J. (2006). Neural mechanisms of addiction: the role of reward-related learning and memory. *Annu Rev Neurosci*, 29:565–598.

Jaakkola, T., Jordan, M., and Singh, S. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Res*, 39(21):3621–3629.

Jazayeri, M. and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nat Neurosci*, 9(5):690–696.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical

and computational perspectives. *Neural Netw*, 15(4-6):535–547.

Kable, J. W. and Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nat Neurosci*, 10(12):1625–1633.

Kaelbling, L. P. (1993). *Learning in Embedded Systems*. MIT Press, Cambridge, MA.

Kakade, S. and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Netw*, 15(4-6):549–559.

Killcross, S. and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb Cortex*, 13(4):400–408.

Körding, K. (2007). Decision theory: what "should" the nervous system do? *Science*, 318(5850):606–610.

Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247.

Krebs, J., Kacelnik, A., and Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature*, 275:27–31.

Lengyel, M. and Dayan, P. (2008). Hippocampal contributions to control: The third way. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA.

Lo, C.-C. and Wang, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nat Neurosci*, 9(7):956–963.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci*, 9(11):1432–1438.

Mangel, M. and Clark, C. W. (1989). *Dynamic Modeling in Behavioral Ecology*. Princeton University Press.

McClure, S., Gilzenrat, M., and Cohen, J. (2006). An exploration–exploitation model based on norepinephrine and dopamine activity. *Advances in neural information processing systems*, 18:867–874.

McClure, S. M., Laibson, D. I., Loewenstein, G., and Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695):503–507.

McNamara, J. and Houston, A. (1980). The application of statistical decision theory to animal

behaviour. *J Theor Biol*, 85(4):673–690.

Montague, P. R. (2006). *Why Choose This Book?: How We Make Decisions*. Dutton Adult.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J Neurosci*, 16(5):1936–1947.

Neal, R. M. and Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants'. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers:.

Ng, A., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337.

Parker, A. J. and Newsome, W. T. (1998). Sense and the single neuron: probing the physiology of perception. *Annu Rev Neurosci*, 21:227–277.

Platt, M. L. and Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238.

Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Pyke, G. (1984). Optimal Foraging Theory: A Critical Review. *Annual Review of Ecology and Systematics*, 15(1):523–575.

Rao, R. P. N. (2004). Bayesian computation in recurrent neural circuits. *Neural Comput*, 16(1):1–38.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*, 20(4):873–922.

Ratcliff, R. and Rouder, J. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, 9(5):347–356.

Redgrave, P., Gurney, K., and Reynolds, J. (2007). What is reinforced by phasic dopamine signals? *Brain Res Rev*.

Rescorla, R. and Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, pages 64–99.

Roitman, J. D. and Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J Neurosci*, 22(21):9475–9489.

Ross, S. (1983). *Introduction to Stochastic Dynamic Programming: Probability and Mathematical*. Academic Press, Inc. Orlando, FL, USA.

Sahani, M. and Dayan, P. (2003). Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Comput*, 15(10):2255–2279.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36(2):241–263.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., and Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–667.

Shadlen, M. N., Britten, K. H., Newsome, W. T., and Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci*, 16(4):1486–1510.

Shadlen, M. N., Hanks, T. D., Churchland, A. K., Kiani, R., and Yang, T. (2007). The speed and accuracy of a simple perceptual decision: A mathematical primer. In Doya, K., Ishii, S., Pouget, A., and Rao, R. P., editors, *Bayesian Brain: Probabilistic Approaches to Neural Coding*, chapter 10, pages 209–238. MIT Press, Cambridge, MA.

Shadlen, M. N. and Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proc Natl Acad Sci U S A*, 93(2):628–633.

Smith, P. L. and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends Neurosci*, 27(3):161–168.

Stocker, A. A. and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci*, 9(4):578–585.

Suri, R. E. and Schultz, W. (1998). Learning of sequential movements by neural network model

with dopamine-like reinforcement signal. *Exp Brain Res*, 121(3):350–354.

Sutton, R. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44.

Sutton, R. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proceedings of the Seventh International Conference on Machine Learning*, 216:224.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press.

Trommershäuser, J., Landy, M. S., and Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychol Sci*, 17(11):981–988.

Trommershäuser, J., Maloney, L. T., and Landy, M. S. (2003a). Statistical decision theory and the selection of rapid, goal-directed movements. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1419–1433.

Trommershäuser, J., Maloney, L. T., and Landy, M. S. (2003b). Statistical decision theory and tradeoffs in the control of motor response. *Spat Vis*, 16(3-4):255–275.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychol Rev*, 108(3):550–592.

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

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):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*. PhD thesis, University of Cambridge.

Whiteley, L. and Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision*, 8(3):1–15.

Wickens, J. (1990). Striatal dopamine in motor activation and reward-mediated learning: steps towards a unifying model. *J Neural Transm Gen Sect*, 80(1):9–31.

Wickens, J. R., Horvitz, J. C., Costa, R. M., and Killcross, S. (2007). Dopaminergic mechanisms in actions and habits. *J Neurosci*, 27(31):8181–8183.

Williams, R. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforce-

ment Learning. *Reinforcement Learning*, 8:229–256.

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

Yang, T. and Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, 447(7148):1075–1080.

Yoshida, W. and Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*, 50(5):781–789.

Yu, A. J. (2007). Optimal change-detection and spiking neurons. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1545–1552. MIT Press, Cambridge, MA.

Yuille, A. and Bülthoff, H. (1996). Bayesian decision theory and psychophysics. In Knill, D. and Richards, W., editors, *Perception as Bayesian inference*, pages 123–161. Cambridge University Press New York, NY, USA.

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Comput*, 10(2):403–430.

List of Figures

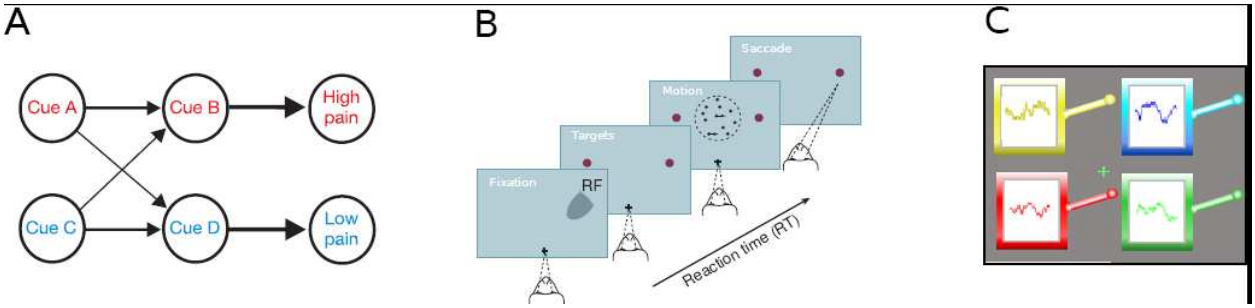

Figure 1: N-DM002

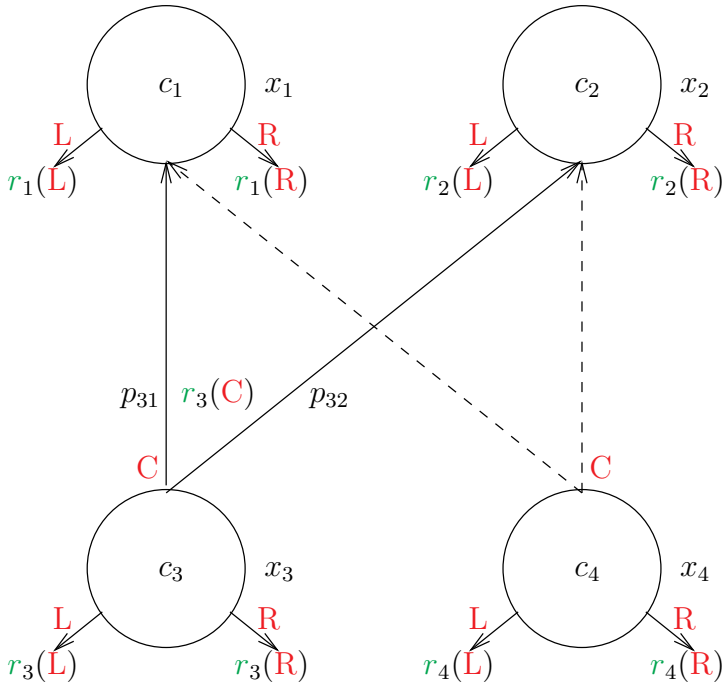

Figure 2: N-DM002

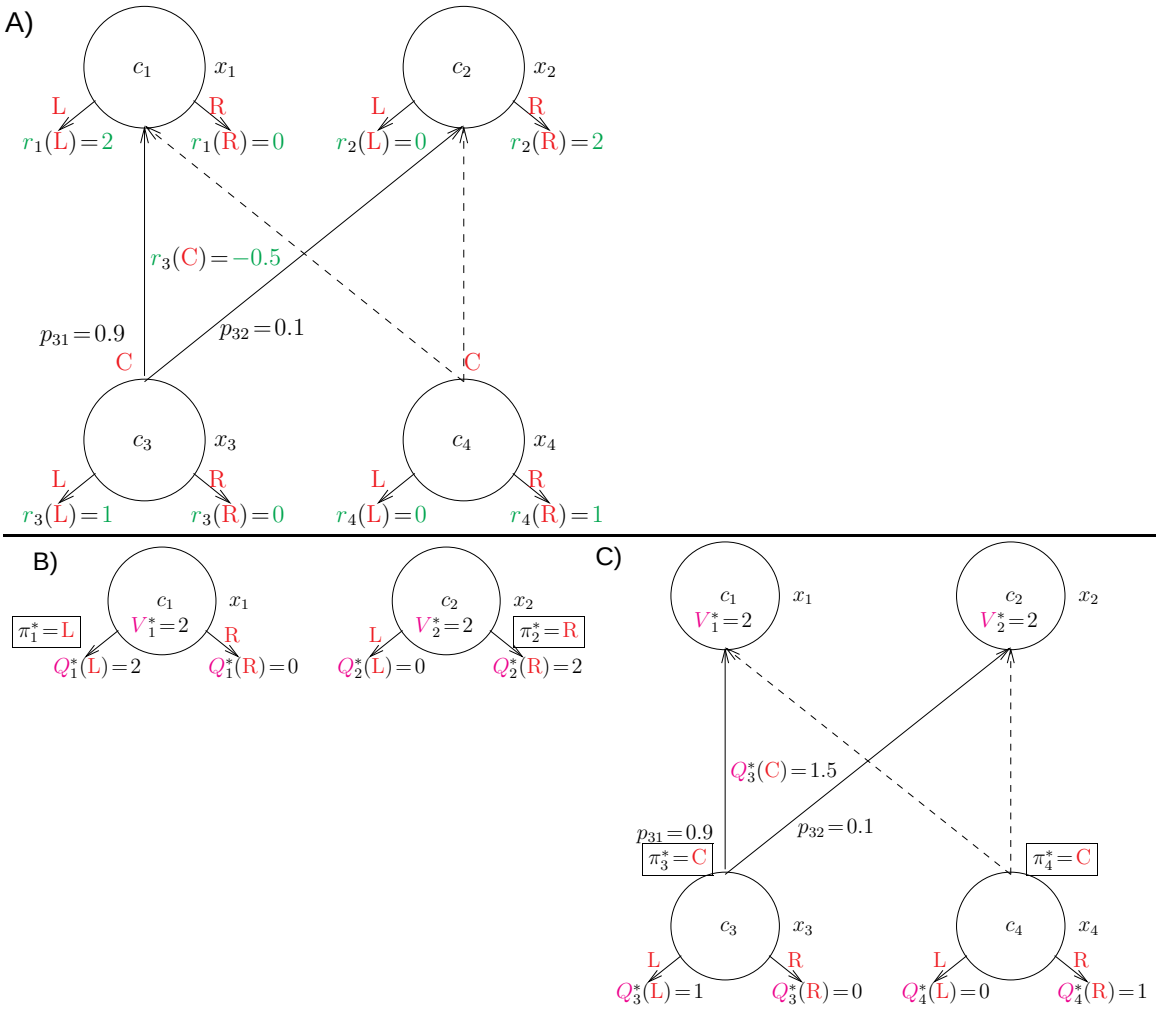

Figure 3: N-DM002

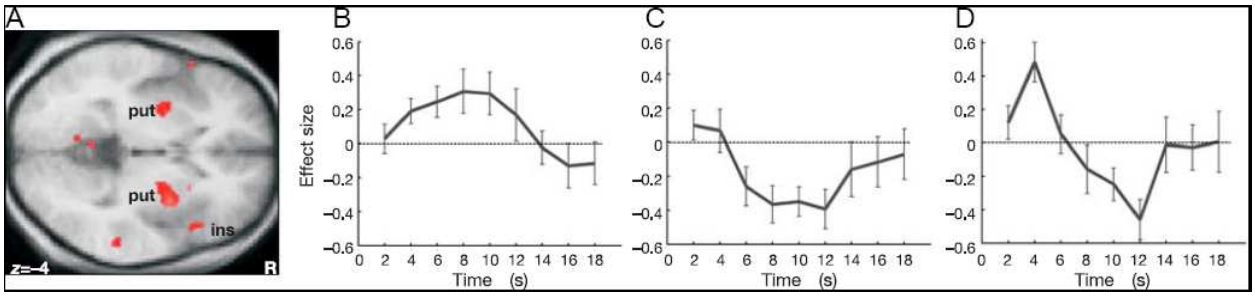

Figure 4: N-DM002

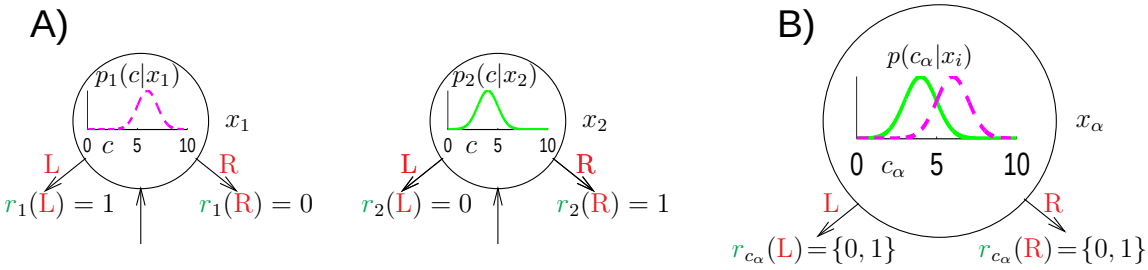

Figure 5: N-DM002

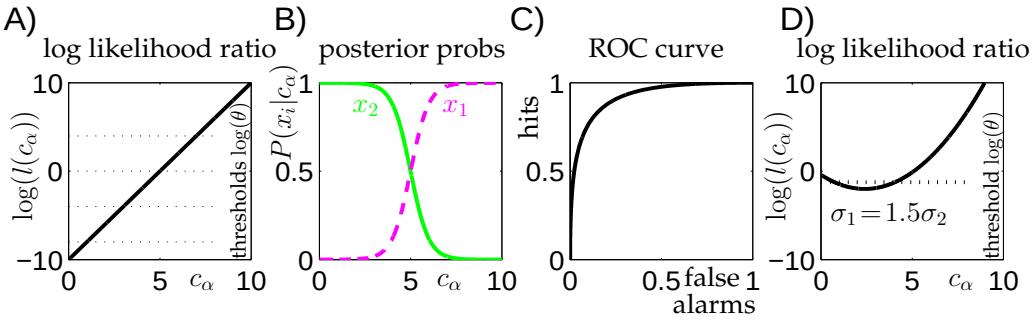

Figure 6: N-DM002

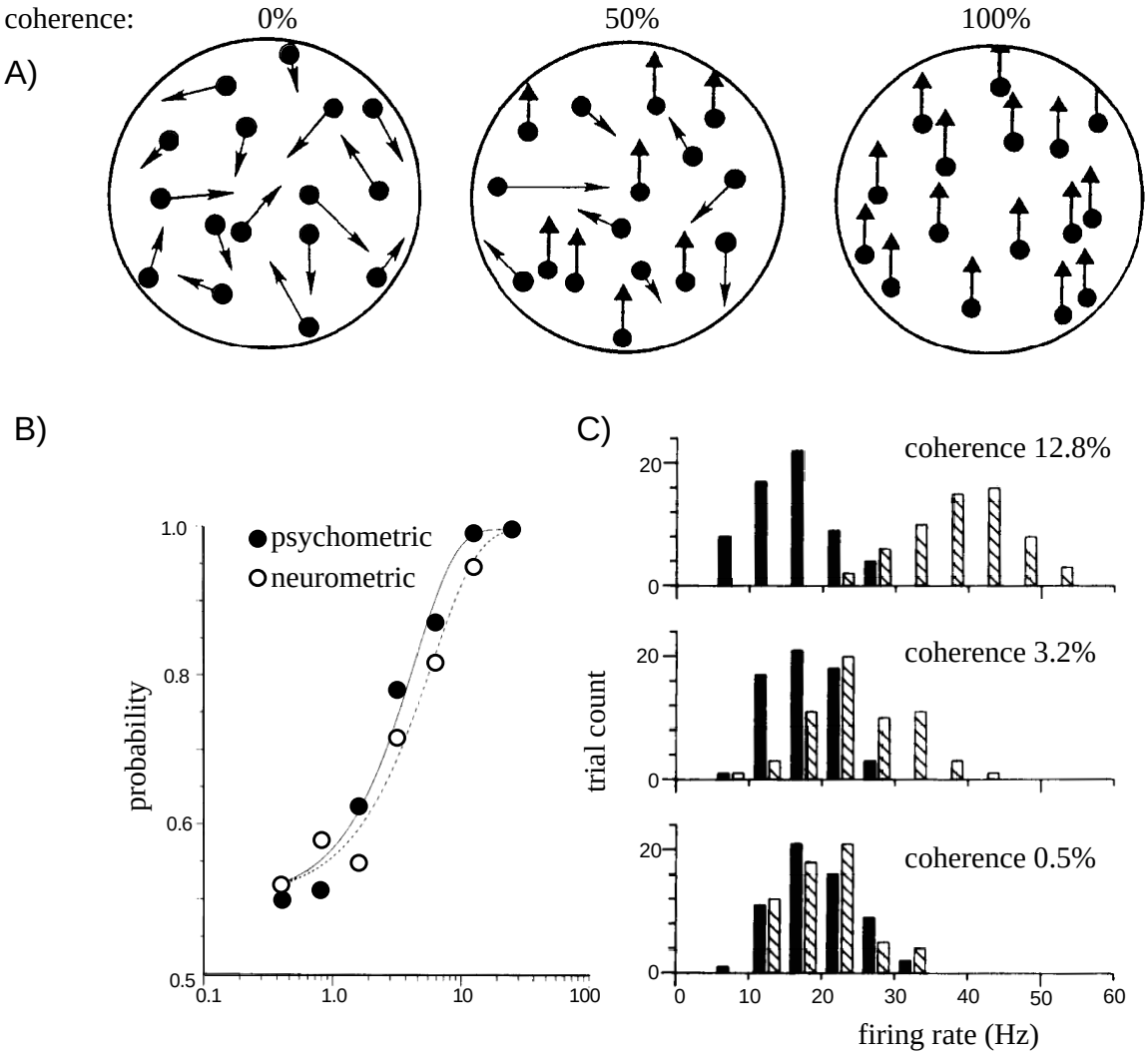

Figure 7: N-DM002

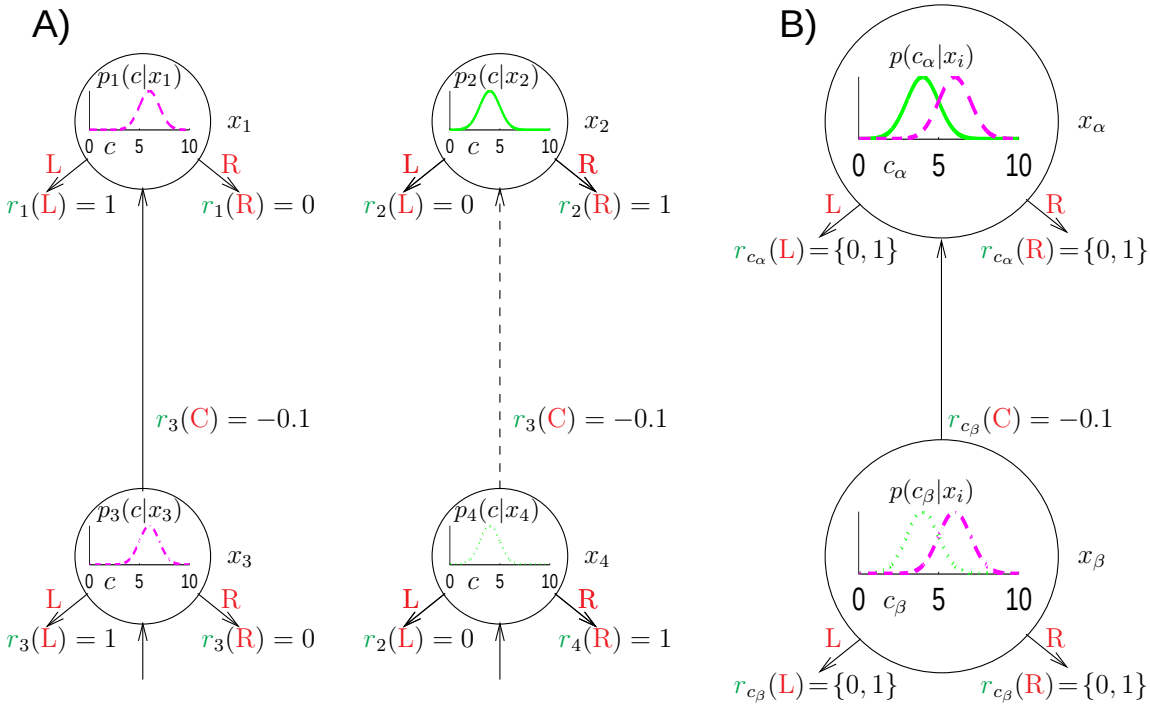

Figure 8: N-DM002

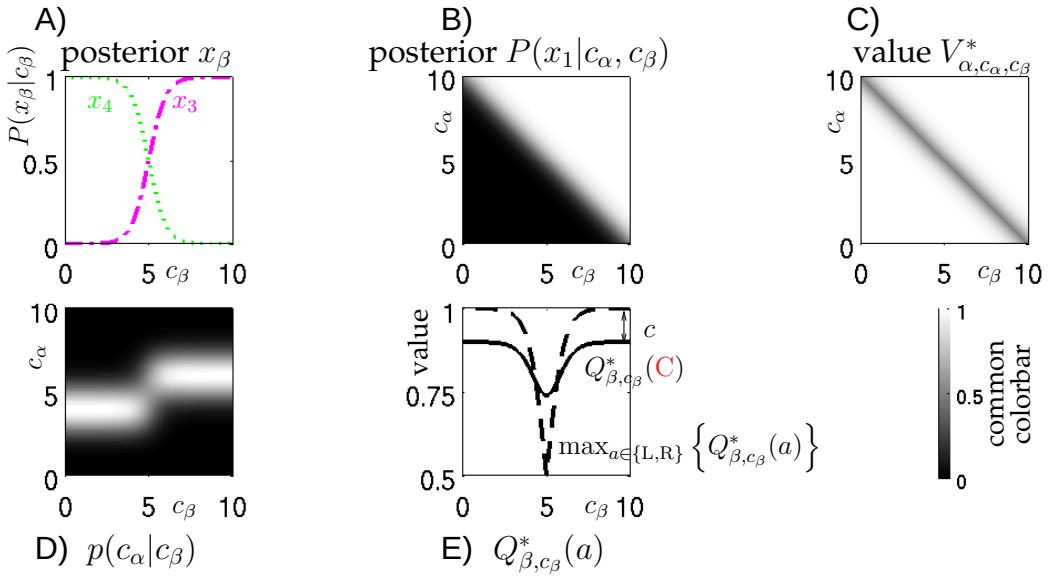

Figure 9: N-DM002

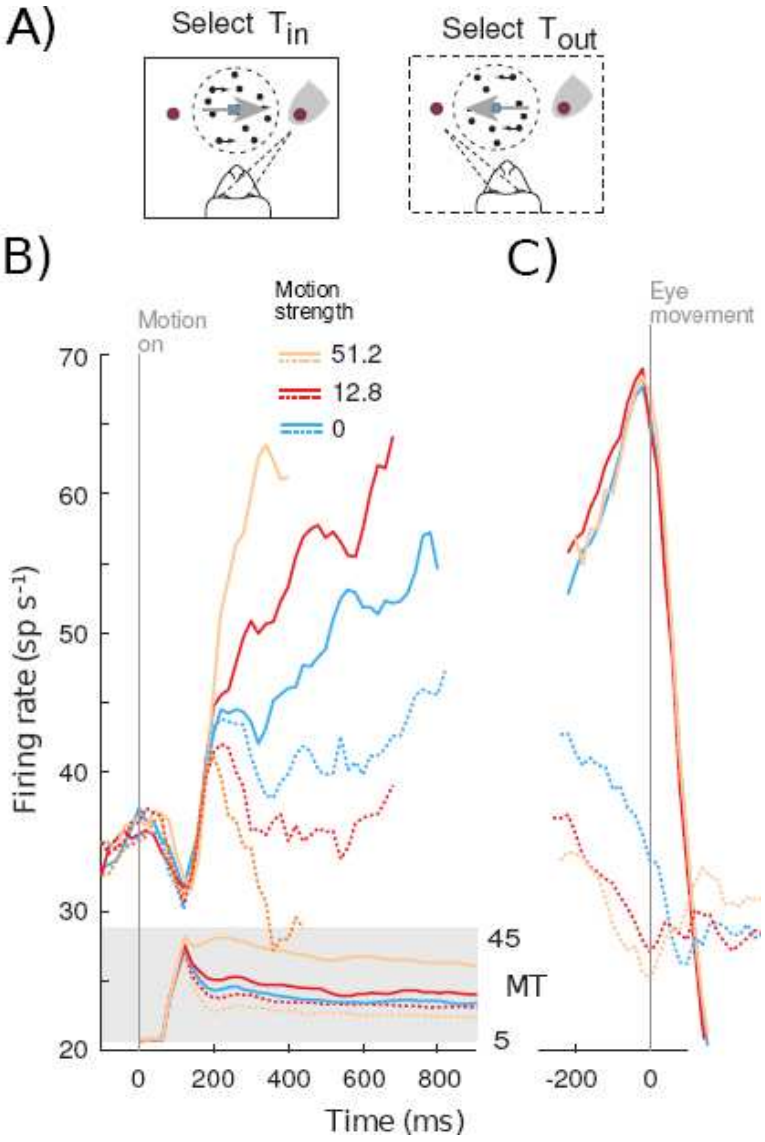

Figure 10: N-DM002

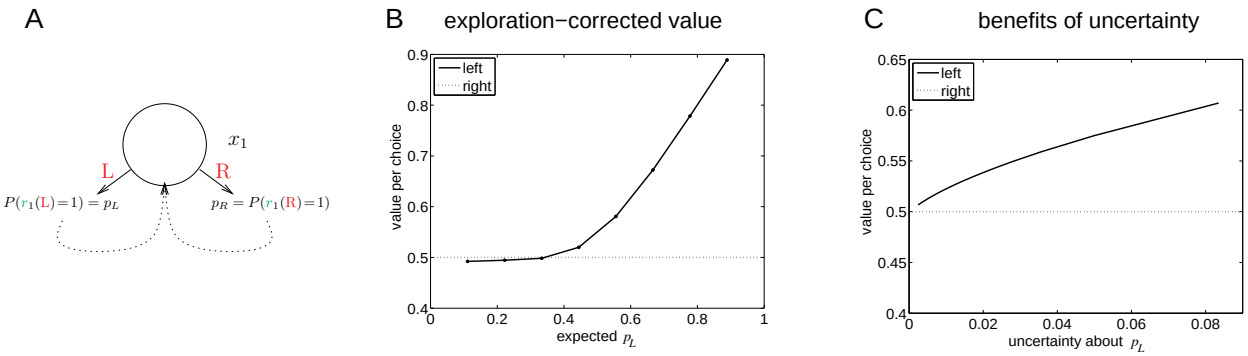

Figure 11: N-DM002

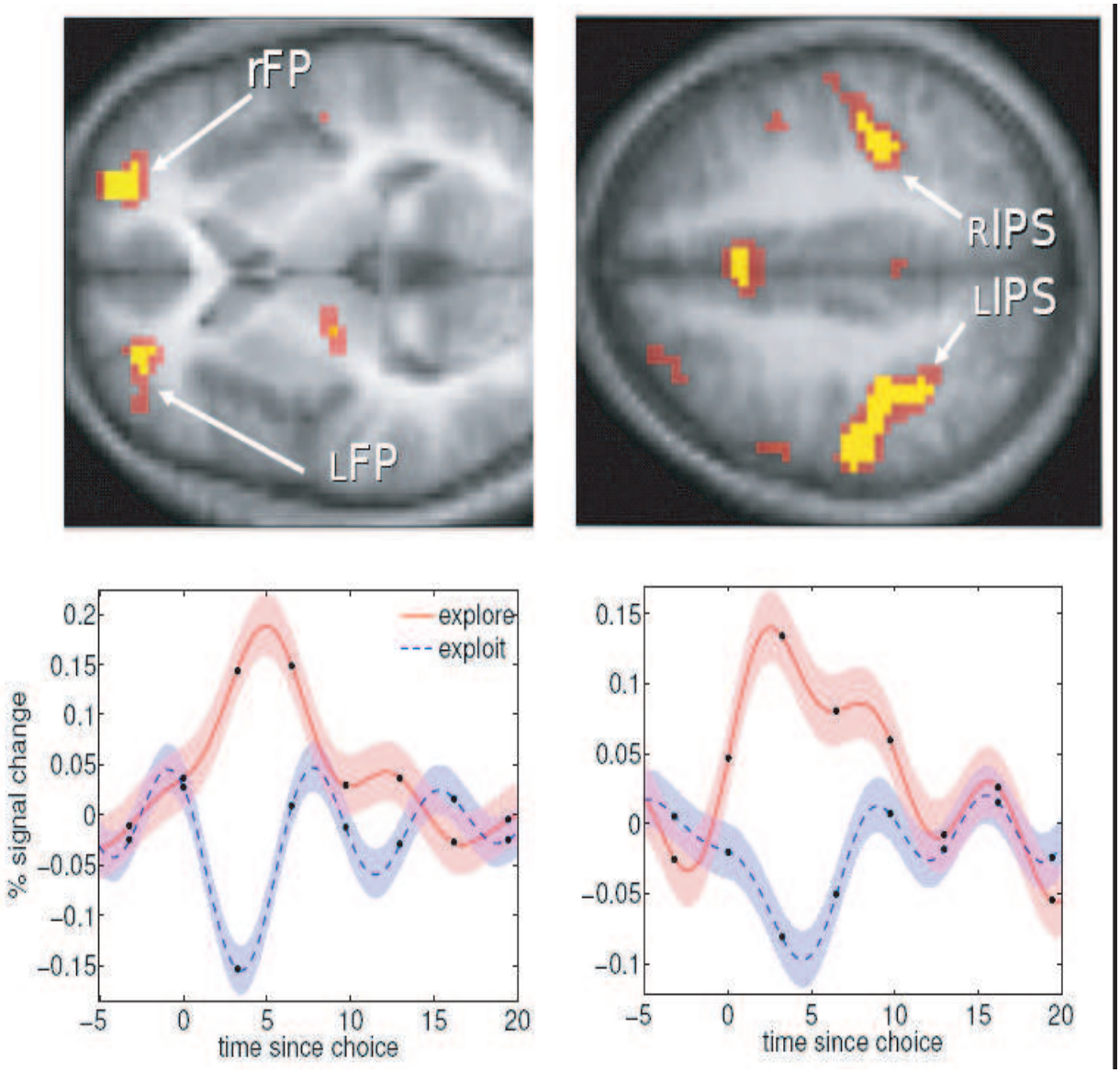

Figure 12: N-DM002